bookboon.com

3D Video Processing and Transmission Fundamentals

Chaminda Hewage

Chaminda Hewage

3D Video Processing and Transmission Fundamentals

3D Video Processing and Transmission Fundamentals
1st edition

ISBN 978-87-403-0810-5

Contents

Abstract

3D video provides the sensation of depth by adding the depth dimension to conventional 2D imagery and video. This allows our human visual system (HVS) to perceive depth as we do in normal vision, which fuses two slightly different views of the same scene in the brain. At present, 3D video applications are not only limited to flight simulators and IMAX theatres, but also available in mobile phones (e.g. LG Optimus), tablets (e.g. GADMEI 3D tablet), television (almost all makes), and advertising boards. Currently, most of the 3D content is user generated and 2D to 3D conversions while percentage of service provider generated content is reducing. This added dimension of depth in 3D imagery and video comes at a cost. Unlike 2D video, the 3D video contents are bulky in nature and often require a larger storage, memory, processing power and bandwidth for communication applications. For instance, the stereoscopic video which is regarded as one of the simplest forms of 3D video, requires twice the space of 2D video since binocular/stereo video content consists of two video streams generated for left and right eyes. This is a major challenge when it comes to delivering 3D video over band-limited channels such as wireless channels. Therefore, it is necessary to have efficient compression and transmission methods to enable 3D video over already established infrastructures for 2D video storage and transmission. This text book presents the methodologies which could be adapted to compress 3D video. Furthermore, this elaborates on effective transmission approaches for 3D video. Perceptual aspects of 3D video technologies also recently received much attention due to the complex nature of 3D perception. Therefore, this book also elaborates on quality evaluation of 3D video. The latest research efforts are also briefly presented to provide a glance of where the technology is heading. The outline of the text book is listed below.

Chapter 1: Introduction

Chapter 2: Stereo vision, 3D video capture and scene representations

Chapter 3: Stereoscopic 3D video compression

Chapter 4: The transmission aspects of 3D video

Chapter 5: 3D video display technologies

Chapter 6: Quality evaluation of 3D video

Chapter 7: Conclusion

Book Description

3D video provides us the sensation of depth by adding a depth dimension to already existing 2D video. This provides the users, improved quality of experience (QoE), natural viewing conditions and a supportive platform for human interaction. On the other hand, 3D video in medical applications (e.g., robotic surgery, minimal invasive surgery (MIS)) could improve the diagnosis and accuracy of surgical procedures. However, the demand for resources (e.g., a large storage and high bandwidth for communication) is hindering the deployment of 3D video applications into a wider market. This book elaborates on the major components of end-to-end 3D video communication chain and discusses the current issues and potential solutions using existing technologies and infrastructures. The main topics covered in this book are different 3D-video formats, 3D video capture technologies, 3D video encoding methods, 3D video transmission approaches, and 3D video quality evaluation aspects.

Author Description

Dr. Chaminda T.E.R. Hewage received the B.Sc. Engineering (Hons.) degree in Electrical and Information Engineering from University of Ruhuna, Sri Lanka. During 2004-2005, he worked as a Telecommunication Engineer for Sri Lanka Telecom in the field of Data Communications. He obtained his Ph.D. (Thesis title: „Perceptual quality driven 3D video over networks“) from Centre for Communication Systems Research (CCSR), University of Surrey, Guildford, UK. At present, he is attached to Wireless Multimedia & Networking Research (WMN) Group at Kingston University-London, UK. His current research interests are 2D/3D video processing and communications, error resilience/concealment, real-time image/video quality evaluation and related fields in multimedia communications.

1 Introduction

Your goals for this "Introduction" chapter are to learn about:

- Recent developments of 3D video.
- Identify the components of 3D video end-to-end chain.
- Major challenges for 3D video application deployment.
- A brief description about contents covered in this book.

Recent developments in audio/video/multimedia capture, real-time media processing capabilities, communication technologies (e.g., Long Term Evaluation (LTE)), and display technologies (e.g., 3D displays) are now facilitating rich multimedia applications beyond conventional 2D video services. 3D video reproduces real-world sceneries as viewed by the human eyes. It provides a state of 'being there' or 'being immersed' feeling to its end users. Moreover, the consumers will be more pleased with immersive video than the computer generated 3D graphics. 3D video is described in technical terms as "geometrically calibrated and temporally synchronized (group of) video data or image-based rendering using video input data" in [1]. According to [1] another possible definition is image-based rendering using video input data or video based rendering. The necessary technologies to realize 3D video services over communication networks are illustrated in Figure 1.1. The technological advancements in 3D video capture, representation, processing, transmission and display will enable the availability of more and more immersive video applications to the consumer market at an affordable cost. This will further improve the comfortness in 3D viewing and quality of experience in general. Therefore, in the future, 3D media applications will not be limited to flight simulators, cyberspace applications and IMAX theatres. 3D video applications will enhance the quality of life in general by capturing home and office media applications (e.g. video conferencing, video broadcasting, broadband video, etc.).

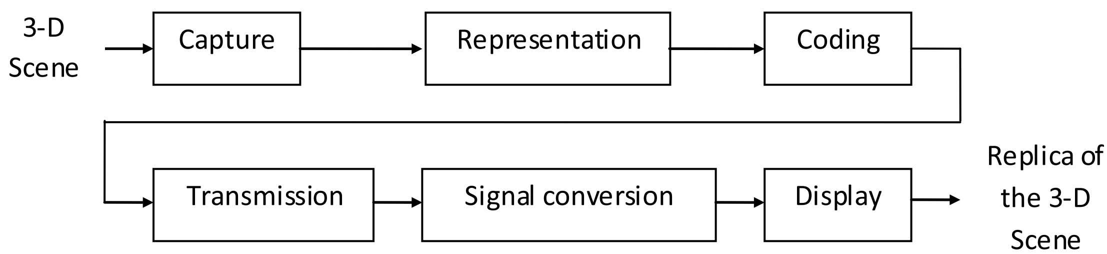


Figure 1.1: 3D video chain

Stereoscopic video is one of the simplest forms of 3D video. It provides the sensation of depth to end users through rendering of two adjacent views of the same scene. Moreover, this 3D video representation has the potential to be the next step forward in the video communication market due to its simple scene representation and adaptability to existing audio-visual technologies. In order to support 3D video services, the existing 2D video application scenarios need to be scaled into a fourth dimension, called "the depth". The availability of multimedia content in 3D will enhance the overall quality of reconstructed visual information. Therefore, this technology will bring us one step closer to the true representation of real-world sceneries. Moreover, 3D video technologies will improve our Quality of Experience (QoE) in general at home and in the work place. The main challenge of these emerging technologies is to adapt them into the existing video communication infrastructure in order to widely disseminate the content during the introduction/migration phase of these new multimedia technologies.

Even though the initial developments of 3D video technologies are in place, there are a several open areas to be investigated through research. For instance, the storage and transport methods (i.e. signaling protocols, network architectures, error recovery) for 3D video are not well exploited. Moreover, the addressing of these problems is complex due to the diversity of different 3D video representations (e.g. stereoscopic video, multi-view video). In addition, the ways and means of fulfilling the extensive demand for system resources (e.g. storage and transmission bandwidth) need to be addressed. Furthermore, the backward compatibility and scalability issues of these applications need to be addressed in order to facilitate the convergence/integration of these services with the existing 2D video applications. The evaluation of 3D video quality is important to quantify the effects of different system parameter settings (e.g. bitrate) on the perceived quality. However, the measurement of 3D video quality is not straight forward as in 2D video due to multi-dimensional perceptual attributes (e.g. presence, depth perception, naturalness, etc.) associated with 3D viewing. Therefore, much more investigation needs to be carried out to simplify the quality evaluation of 3D video or 3D QoE. This book has presented the proposed solutions for some of the issues mentioned above with a major focus on 3D video compression and transmission, which are described below.

The captured 3D video content is significantly larger than 2D video content. For example, stereoscopic video could be twice the size of a conventional 2D video stream, as it has two closely related camera views. As a result, 3D video requires a large storage capacity and high transmission bitrates. In order to reduce the storage and bandwidth requirements, the immersive video content needs to be efficiently compressed. Existing video compression algorithms may or may not be suitable for encoding 3D video content. Moreover, the unique characteristics of 3D video can be exploited during compression in order to further reduce the storage and bitrate required for these applications. The transmission of these contents should be easily synchronized among different views during playback. In addition, backward compatibility with conventional 2D video applications would be an added advantage for emerging 3D video applications.

Transmission of 3D content is also a major challenge due to the larger size of the 3D video content. Therefore, effective mechanisms need to be in place to compress 3D video content into a more manageable size to be transmitted over band-limited communication channels. On the other hand, the transmission of immersive video content could be optimized based on the perceptual importance of the content. For instance, the different elements of the 3D video content can be prioritized over communication channels based on their error sensitivities. These prioritized data transmission schemes can be effectively used in optimizing the resource allocation and protection for immersive media content over error prone communication channels without any degradation to the perceived quality of the reconstructed 3D replica. The quality of transmitted video suffers from data losses when transmitted over an error prone channel such as wireless links. This problem is also common for emerging 3D video communication applications. The effect of transmission errors on perceived 3D quality is diverse in nature due to the multi-dimensional perceptual attributes associated with 3D viewing. Therefore, efficient error resilient and error concealment algorithms need to be deployed to overcome the detrimental effects that occur during transmission. Existing error recovery techniques for 2D video could also be used in recovering corrupted frames. Moreover, error resilient/concealment techniques which are tailor-made to particular types of 3D video could be implemented at the application level.

This book investigates and presents efficient 3D compression and transmission technologies which offer improved compression efficiency, backward compatibility, efficient error recovery and perceptually prioritized data transmission. Even though 3D video comes in different scene representations (e.g. Omni-directional video and Multi-view video), this book focuses on facilitating stereoscopic video communications, since stereoscopic video has the potential to be easily adopted into the existing video communication infrastructure compared to other complex representations of 3D video. The first chapter provides the rationale and a brief description of the book while the final chapter, Chapter 7, summarizes the 3D video concepts covered in this book and discusses the potential areas for future research in efficient and robust 3D video communications. The work presented in the other chapters is summarized below.

Chapter 2 describes stereo vision, the state of the art 3D video technologies for scene capture and different scene representations of 3D video. Then, existing multimedia compression technologies are described with more specific details about 3D video coding techniques in Chapter 3. In Chapter 4, the transmission aspects of 3D video and potential application scenarios are presented. Furthermore, an introduction to error resilience and error concealment techniques used in multimedia communication is presented. The display technologies and viewing aids associated with potential 3D video applications are also discussed in Chapter 5. Finally, an explanation of measuring 3D video quality subjectively and objectively is presented in Chapter 6.

2 Stereo vision, 3D video capture and scene representations

Your goals for this "Introduction" chapter are to learn about:

- Different 3D video representations.
- Stereoscopic video capture technologies.
- 3D Image Warping/Depth Image Based Rendering (DIBR).
- The difference between left and right views based stereoscopic video vs. colour plus depth 3D video.

2.1 Different 3D video representations

3D objects can be reconstructed from the captured real world images, which provide the user the impression of 3D video. The methods of reconstruction and capture of the image sequences are based on the requirements of the targeted application scenario. According to the classification of MPEG-3DAV (Motion Picture Expert Group-3D Audio Visual), three scene representations of 3D video have been identified, namely omni-directional (panoramic) video, interactive multiple-view video (free-viewpoint video) and stereo video [2]. Omni-directional video allows the user to look around a scene (e.g. IMAX-Dome). This is an extension of planar 2D image into a spherical or cylindrical image plane. Figure 2.1 shows some example omni-directional images generated with the Dodeca™ 1000 camera system and post-processed with corresponding Immersive Media technology [3].

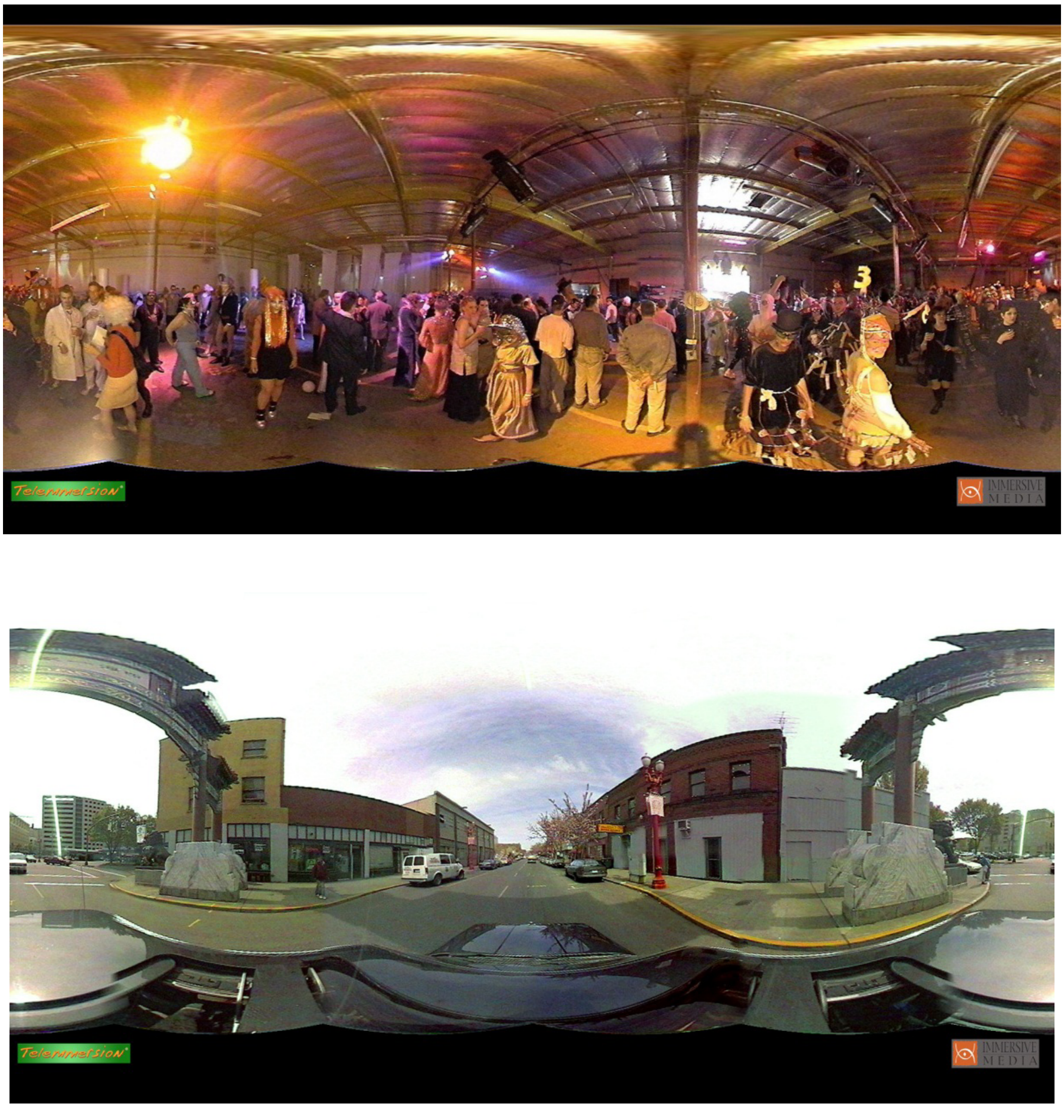
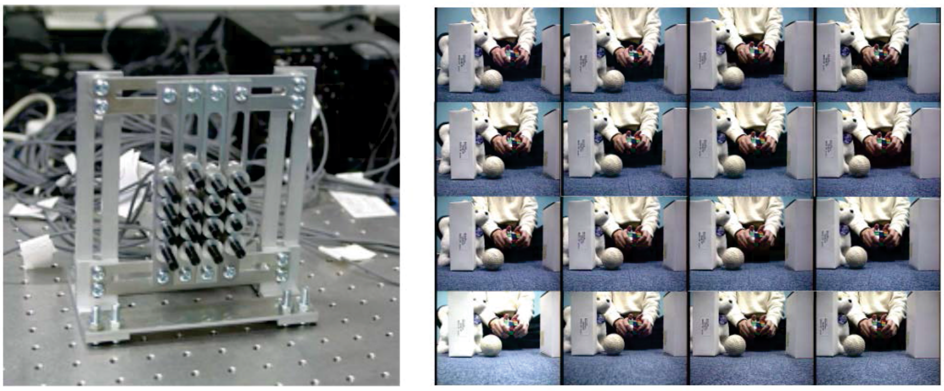


Figure 2.1: Omni-directional images from Telemmersion® video

The potential applications scenarios relevant to Omni-directional video are described in [4]. However, the 3D video in this format has a limited application scope (e.g. navigation and object avoidance) and may not be suitable for general application scenarios like 3D TV broadcasting. Multi-view video (e.g. free-viewpoint video) is the general case for all 3D scene representations. It allows the user to select an arbitrary viewpoint and direction within a visual scene, generating an immersive environment. It generates virtual camera views through interpolation of real camera views. This representation can be effectively utilized in wide range of applications, including FTV (free-viewpoint television) and surveillance [2]. The ray space approach and 3D model based approaches have been identified for real-time rendering of novel views [5]. Figure 2.2 shows an array of cameras (i.e. 16 × 16) which can capture multiple raw video sequences and the captured multiple videos [6]. However, due to the high demand for system resources (e.g. processing power, bandwidth, and storage), the availability of multi-view video applications to the mass market will be further delayed till 3D video technologies and supporting infrastructure get to a more mature stage than the current stage of the development process. The third approach is stereoscopic video which we describe in more details in the next section.



(a) (b)

Figure 2.2: Multi-view camera platform; (a) Input system, (b) Captured images [6]

2.2 Stereoscopic video and capture technologies

In order to produce a 3D impression, stereoscopic video representation renders two adjacent views for each eye of the user. The left and right views are then fused in the visual cortex of the brain to perceive the depth of a scene (see Figure 2.3). According to [7], accurate perception of depth by human visual system can be attributed to two main categories of depth cues. The physiological and psychological depth cues mentioned in [7] are as follows:

Physiological depth cues

- Binocular disparity: The dissimilarity in views due to the relative location of each eye.
- Accommodation: The change in the focal length of the lens in the eye caused by muscles in the eye to produce a focused image on the retina.
- Convergence: The rotation of eyes to align or merge the left and right eye images into a single image with perceived depth.
- Motion parallax: The difference in views produced by moving the scene or the viewer. For example, in a movie it is possible to realize the size of an object which is speeding towards the viewer based on the relative change in size with time. This cue often differentiates the realism of a video from that of a still image.
- Chroma-stereopsis: The difference in apparent depth due to the colour of an object from refraction effects in the eyes.

Psychological depth cues

- Image size: This is a useful hint but not sufficient to determine size or depth of objects.
- Linear perspective: This is the decrease in the apparent size of an object with increasing distance.
- Aerial perspective: This refers to the hazy and bluish appearance of distant objects.
- Shading suggests that objects farther from the source of light are darker.
- Shadowing of an object on others provides clues about position and size.
- Occlusion: of objects provides a clue about their relative location.
- Texture gradient provides clues regarding distance and relative location.
- Brightness of an object suggests that it is closer than dimmer objects.

Stereoscopic video capturing system mainly exploits the binocular disparity cue which helps human visual system to perceive depth.

Stereoscopic video is one of the simplest forms of 3D video available in the literature. Moreover, this representation is a subset of multiple-view video, where only two camera views are rendered to the user. Due to the simple representation and adaptability (e.g. simple camera arrangement, cost effective display systems, etc), stereoscopic video could be employed in future broadcasting (e.g. 3D TV), storage and communication applications (e.g. 3D conferencing) relatively easily compared to other representations of 3D video. The existing infrastructure for audio-visual technologies (e.g. compression/decompression) can be adopted to send binocular content over communication channels. Moreover, the demand for resources (e.g. bandwidth and processing power) will be lower compared to the multi-view video. Therefore, in this book, stereoscopic video is considered as the main 3D video representation and the constraints and problems associated with stereoscopic video communications are presented.

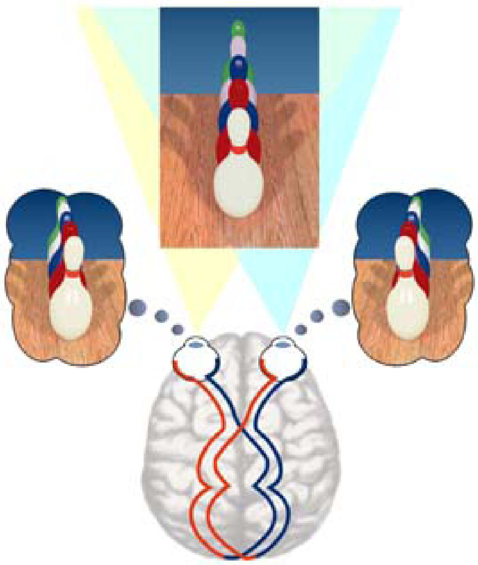


Figure 2.3: Illustration of stereo vision [8]

At present, technologists are working on several 3D video applications, which cover the whole application chain consisting of 3D capture, compression, transmission, rendering of 3D video and high-end display technologies [9]. The separate modules in a stereoscopic video chain are shown in Figure 2.4. The term “cyclopean view” in Figure 2.4 refers to the intermediate view between the left and right view perspectives. In order to support stereoscopic video for single-user or multi-user display systems all the interconnected technologies (capture, transmission) should work in harmony. More importantly, the sufficient stereoscopic video content should be generated to meet the customer demand.

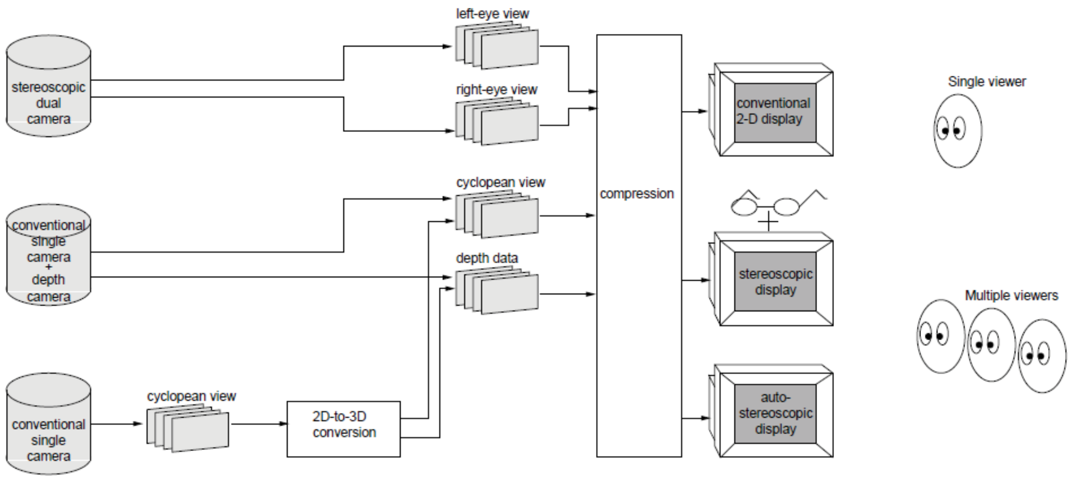


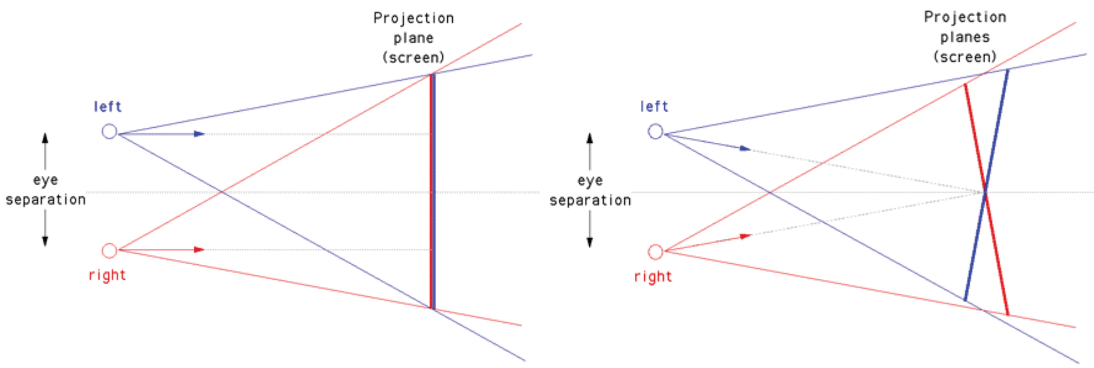
Figure 2.4: Separate modules in stereoscopic video chain [10]

There are several techniques to generate stereoscopic video material including dual camera configuration, 2D-to-3D conversion algorithms, 3D/Depth-range cameras [10]. Stereoscopic view of a scene captured using a stereo camera pair (i.e. the left-eye and the right-eye view are recorded separately by two cameras taken from a slightly different perspective) is the simplest and most cost effective way to capture stereo video at the moment compared to other technologies available in the literature (see Figure 2.5). The shooting parameters such as camera base distance (distance between the two cameras), convergence distance (distance of the cameras to the point where both optical axis intersect) and camera lens focal length can be utilized to scale the horizontal disparity and thus the degree of perceived depth. Furthermore, 3D video with the dual camera configuration provides fewer burdens at the receiver side for rendering 3D video due to the availability of two views. Two dual camera configurations can be distinguished, namely the parallel camera configuration and the toed-in camera configuration, also called converging cameras (see Figure 2.6). According to the study carried out in [11], the parallel camera configuration avoids geometrical distortions like the keystone distortion and depth plane curvature. As dual camera configuration generates two separate image sequences for left and right view, more system resources are necessary to process, store and transmit the generated content in comparison to the resource requirements of 2D video. For example, a double disk space is needed to store the raw left and right video sequences. Moreover, the viewing angle will be limited with the stereo camera pair and thus no interactivity can be employed. The commercially available stereoscopic video cameras and 3D add-ons for standard camcorders are listed in [12].

(a)

(b)

Figure 2.5: Stereo video capture, (a) stereo video camera, (b) Captured scenes



(a) **(b)**

Figure 2.6: Dual camera configurations; (a) Parallel cameras, (b) Toed-in cameras

The 2D-to-3D conversion algorithms can be employed to transform 2D video into 3D video sequences. For instance, existing movies can be viewed as a novel stereoscopic film [13] [14]. In principle, 2D-to-3D conversion algorithms derive a depth map sequence from a 2D still image sequence. According to [15], the depth estimation techniques such as depth from motion and structure from motion will convert only a limited amount of the monoscopic video into 3D video automatically. Therefore, novel 2D-to-3D conversion methods are necessary with a limited manual intervention in order to support off-line and real-time conversion of 2D video into 3D video. The semi-automatic methods/algorithms developed by Dynamic Digital Depth Research Pty Ltd and Philips to recover the depth map of a monoscopic video are presented in [15] and [16] respectively.

The latest addition to the 3D capturing technology is the depth/range cameras. They simultaneously capture a colour image sequence and associated per-pixel depth image sequences of a scene. The 3D camera utilizes a light pulse to measure the relative depth of the objects in the scene (see Figure 2.7 (a)). Figure 2.7 (b) shows the internal architecture of the High Definition (HD) three-dimensional camera developed by NHK Laboratories Japan [17]. The Zcam™ [18] and Axi-vision [19] 3D cameras are two commercially available 3D depth/range cameras, which are developed by 3DV systems and NHK respectively. Moreover, these products are also available as add-ons for existing video capturing devices.

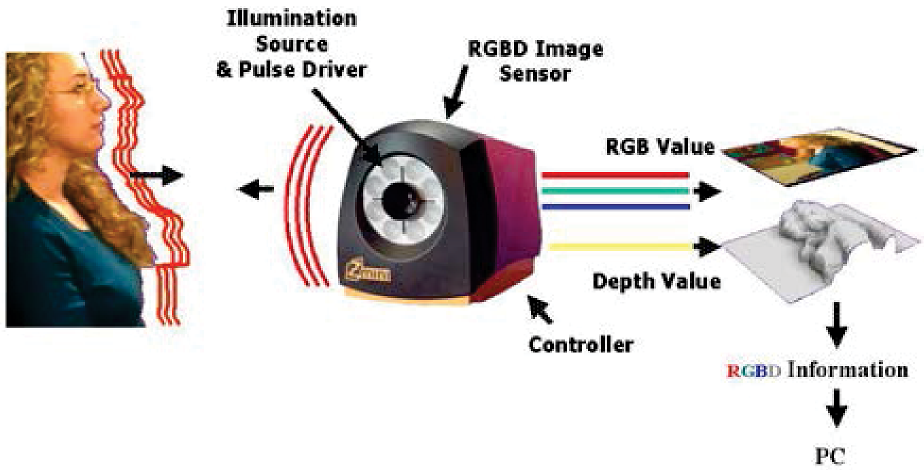
The snapshot of a scene captured with a 3D camera is given in Figure 2.8. The depth map sequence has similar spatio-temporal resolution as the colour image sequence. The depth images can be stored in 8 bit gray values, where gray value 0 specifies the furthest value (i.e. away from camera) and the gray level 255 specifies the closest value (i.e. closer to the camera). In order to translate this depth data representation to real, metric depth values and to support different image sequences with different depth characteristics, the gray levels are normalized into two main depth clipping plains namely;

- The near clipping plane Z_{near} (gray level 255), the smallest metric depth value Z
- The far clipping plane Z_{far} (gray level 0), the largest metric depth value Z.

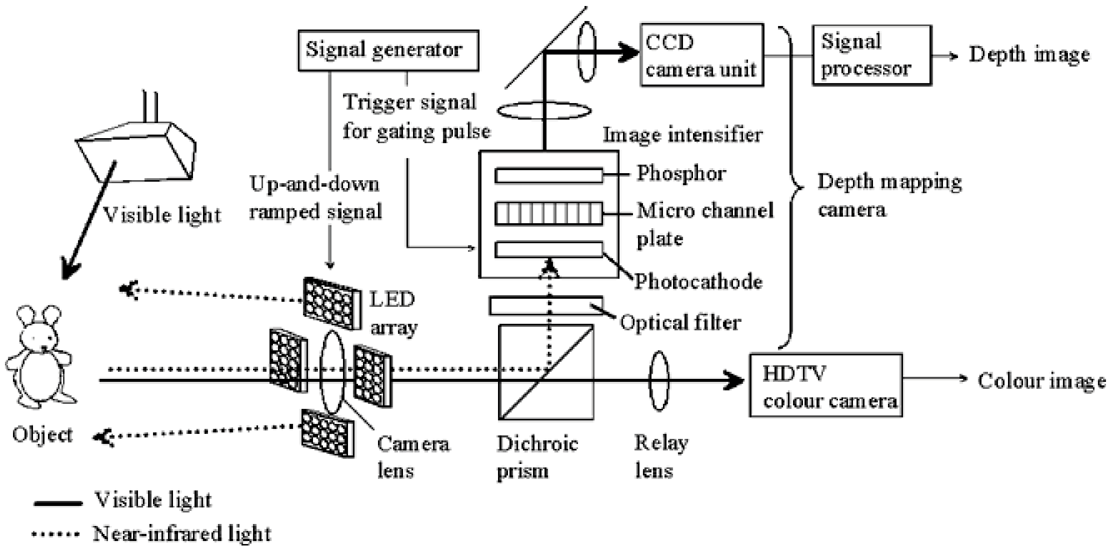
In case of linear quantization of depth, the intermediate depth values can be calculated using Equation 2.1.

$$Z = Z_{far} + v\left(\frac{Z_{near} - Z_{far}}{255}\right) \text{ with } v \in [0,\ldots,255] \qquad \textbf{Equation 2.1}$$

where v specifies the respective gray level value.



(a)



(b)

Figure 2.7: Depth/range camera; (a) Illustration of depth capture, (b) Internal architecture of a 3D camera [17]

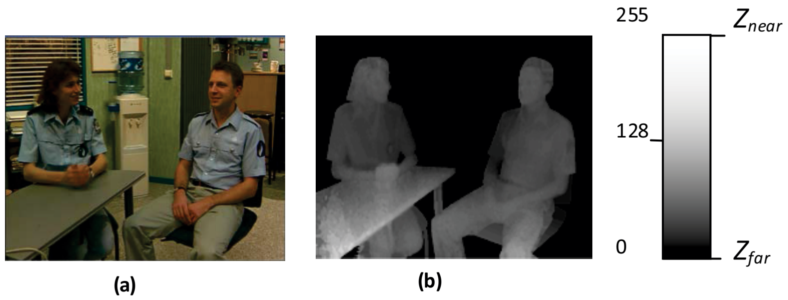


Figure 2.8: Interview sequence; (a) Colour image, (b) Per-pixel depth image. The depth images are normalized to a near clipping plane Z_{near} and a far clipping plane Z_{far}.

Depth-Image-Based Rendering (DIBR) can be utilized to render/synthesize two virtual views for the left and right eyes using the colour image sequence and the corresponding per-pixel depth information [20] [21]. This process can be employed in two major steps;

- The reprojection of original image point into a 3D space with the help of depth information
- The 3D space points are then projected into the image plane of the virtual camera.

In Computer Graphics (CG) this concept is known as 3D image warping. This concept is mathematically derived in subsection 2.2.1.1.

2.3 3D Image Warping

Figure 2.9 shows a system of two cameras and an arbitrary 3D space point X with the projection x and x` in the first and second virtual views respectively. All the image points are on the image plane of π.

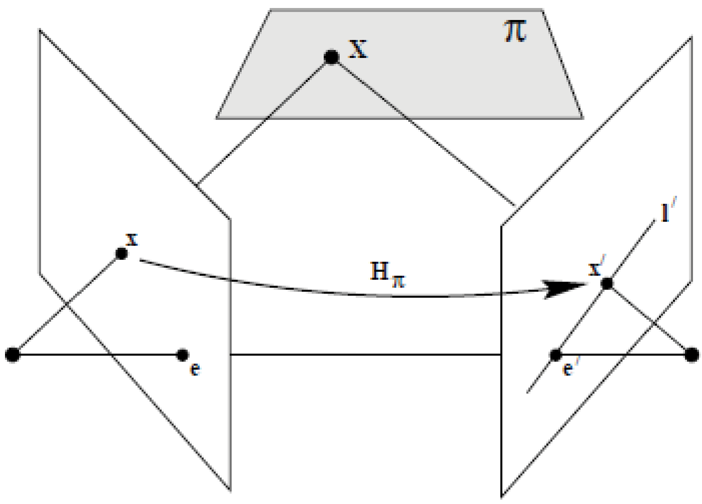


Figure 2.9: A point x in one image is transferred via the plane π to a matching point x` in the second image. The epipolar line through x` is obtained by joining x` to the epipole e`. In symbols one may write $x` = H_\pi x$ and $l` = [e`]\times x` = [e`]\times H_\pi x = Fx$ where $F = [e`]\times H_\pi$ is the fundamental matrix [22].

Under the assumption that the world coordinate system equals the camera coordinate system of the first camera, the two perspective projection equations will be;

$$\tilde{x} \cong AP_n\tilde{X}$$ **Equation 2.2**

$$\tilde{x}' \cong A'P_nH_f\tilde{X}$$ **Equation 2.3**

Where x and x` symbolize two 2D image points with respect to the 3D space point X in homogeneous notation. The symbol ≅ denotes the 'equality up to a non-zero scale-factor' [22] [23]. The 4 × 4 matrix H_f contains the transform matrix which converts the 3D space point from world coordinate system into the camera coordinate system of the second view. The H_f consists of two transform components namely rotation R and translation T. The 3 × 3 matrices A and A` defines the intrinsic parameters of the first and second cameras respectively. The normalized perspective projection matrix is denoted by the 3 × 4 identity matrix P_n.

The 3D space point *X* is still dependent on its depth value *Z*. Hence, Equation 2.2 can be rearranged into;

$$X = ZA^{-1}\tilde{m}$$ **Equation 2.4**

The depth dependent relationship between corresponding points in two perspective views of the 3D scene can be derived using Equations 2.3 and 2.4 and the outcome represents the classical disparity equation (see Equation 2.4).

$$Z'\tilde{m}' = ZA'RA^{-1}\tilde{m} + A't$$ **Equation 2.5**

This 3D image warping relationship can be utilized to render arbitrary novel views with respect to a known reference image. This requires the virtual camera position and orientation relative to the reference camera to be known with the intrinsic parameters of the virtual camera. Then if the depth values of the corresponding 3D space points are known for every pixel in the original image, novel views can be generated using Equation 2.5.

The virtual stereoscopic images can be generated through simplifying the 3D image warping technique (Equation 2.5) to represent horizontal parallax of two virtual camera positions. The relationship will be derived based on the two stereo camera configurations as shown in Figure 2.10. The both configurations can be distinguished based on how they achieve Zero-Parallax Setting (*ZPS*).

- Toed-in: The point of convergence at Z_c achieves through inward-rotation of the left and right cameras.
- Shift-sensor/parallel camera setup: A plane of convergence at Z_c is established through a shift *h* of the camera's CCD (Charged-Couple Device) sensor.

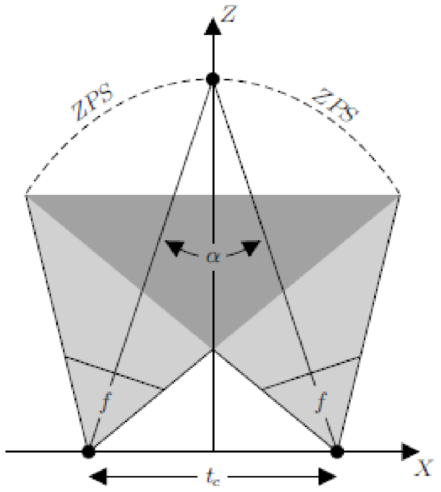


Figure 2.10: Different stereoscopic camera setups a). In the "toed-in" camera setup, a point of convergence at Z_c is established by a joint inward-rotation of the two cameras, b). In the shift-sensor camera setup, a plane of convergence at Z_c is established by a shift h of the camera's CCD sensors. The t_c refers to the inter-axial distance between two cameras [24].

The parallel camera setup is more suitable to be used with the DIBR technique, because all the signal processing steps going to be one-dimensional. With respect to the original view, the virtual cameras (i.e. left and right camera) are symmetrically displaced and their CCD sensors are shifted relative to the position of virtual camera lenses. This sensor shift can be mathematically formulated as a displacement of a camera's principal point *c* [23]. Therefore, intrinsic parameters of the virtual cameras are considered to be having similar intrinsic parameters of the original camera except the horizontal shift h of the respective principal point. This also can be formulated into an equation as follows;

$$A^* = \begin{bmatrix} \alpha_u & 0 & u_0 + h \\ 0 & \alpha_v & v_0 \\ 0 & 0 & 1 \end{bmatrix} = A + \begin{bmatrix} 0 & 0 & h \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \qquad \textbf{Equation 2.6}$$

A^* denotes the intrinsic parameters of either left (i.e. A\`) or right (i.e. A") virtual cameras.

With the assumption that the movement of the two virtual cameras are translational with respect to the reference camera (i.e. Rotation R = I, where I is the 3 × 3 identity matrix) the 3D warping Equation 2.5 can be further simplified using the relationship derived in Equation 2.6;

$$A^* R A^{-1} = A^* A^{-1} = I + \begin{bmatrix} 0 & 0 & h \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \qquad \textbf{Equation 2.7}$$

Substituting Equation 2.5 with the simplified expression in Equation 2.7 the 3D warping equation can be written as;

$$Z^* \widetilde{m}^* = Z \left(\widetilde{m} + \begin{bmatrix} h \\ 0 \\ 0 \end{bmatrix} \right) + A^* t \qquad \textbf{Equation 2.8}$$

With $t_z = 0$, the depth value of the 3D space point is same in the original camera (camera coordinate system of the original view) and the virtual camera (coordinate system of the virtual camera). Therefore, $Z^* = Z$ and Equation 2.8 can be further reduced to;

$$\widetilde{m}^* = \widetilde{m} + \frac{A^* t}{Z} + \begin{bmatrix} h \\ 0 \\ 0 \end{bmatrix} \text{ with } t = \begin{bmatrix} t_x \\ 0 \\ 0 \end{bmatrix} \qquad \textbf{Equation 2.9}$$

Then, the affine pixel positions (u,v) of each warped image can be calculated as;

$$u^* = u + \Delta u \text{ and } v^* = v \qquad \textbf{Equation 2.10}$$

$$= u + \frac{\alpha_u t_x}{Z} + h$$

The horizontal translational distance t_x is equal to the half of the inter-axial distance t_c (i.e. the average eye separation of humans, approximately 64 mm). The translational distance with the direction of the movement is;

$$t_x = \begin{cases} -\frac{t_c}{2} & \text{: Left-eye view} \\ +\frac{t_c}{2} & \text{: Right-eye view} \end{cases}$$ **Equation 2.11**

The amount of sensor shift h is dependent on the chosen convergence distance Z_c. When $Z = Z_c$ the horizontal component u* of the simplified 3D warping Equation 2.10 is identical for both left and right views, i.e. $u` = u”$. Therefore, Equation 2.10 can be rewritten as;

$$h = -t_x \frac{\alpha_u}{Z_c}$$ **Equation 2.12**

where t_x is also defined by Equation 2.11.

Equations 2.10 and 2.11 can be utilized to render the virtual camera views of the parallel stereoscopic camera setup. The characteristics of the rendered virtual left and right views going to be affected by the choice of inter-axial distance t_c, the focal length f of the reference camera and the convergence distance Z_c. Table 2.1 shows how the 3D perception is affected due to these parameter settings. These effects can be attributed to the effects of real stereoscopic camera setup with the adjustments to their camera positions (e.g. inter-axial distance) and intrinsic parameters (e.g. focal length).

Parameter	+/–	Screen parallax	Perceived depth	Object size
Interaxial distance t_c	+	Increase	Increase	Constant
	–	Decrease	Decrease	Constant
Focal length f	+	Increase	Increase	Increase
	–	Decrease	Decrease	Decrease
Convergence distance Z_c	+	Decrease	Shift (forward)	Constant
	–	Increase	Shift (backward)	Constant

Table 2.1: Effects of different stereo camera setup parameters. Qualitative changes in screen parallax values, perceived depth and object size when varying the inter-axial distance t_c, the focal length f or the convergence distance Z_c of a “real” or “virtual” shift-sensor stereo camera setup [25].

2.3.1 The Advantages and Disadvantages of Colour Plus Depth Map Representation

The advantages of using colour plus depth map representation of stereoscopic video compared to the video generated with a stereo camera pair can be listed as follows:

- The 3D reproduction can be adjusted to different stereoscopic displays (e.g. auto-stereoscopic displays) and projection systems as the rendering happens at the receiver side.
- The 2D-to-3D conversion algorithms will generate more colour plus depth stereoscopic video and increase the timely availability of exiting stereoscopic materials.
- Head-Motion Parallax (HMP) could be supported which provides an additional stereoscopic depth cue. This format also limits the viewing angle of the stereoscopic video camera setup.
- Due to the smoothness characteristics of the real world objects the per-pixel depth information doesn't have high frequency components. Thus, the depth sequence can be efficiently compressed with existing compression standards [24] and will require only a limited space and bandwidth compared to the requirements of colour image sequence.
- The diminution of stereoscopic video sensation due to photometrical asymmetries (e.g. in terms of brightness, contrast or colour, between the left and the right eye) will be eliminated as this representation renders the virtual stereo views using the same colour image sequence.
- The depth reproduction can be adjusted at the receiver side based on user preferences (e.g. age, eye strain).
- This representation can be effectively used in 3D post production (e.g. augmenting external objects to the scene using object segmentation with the help of depth information).

However, the existing drawbacks of this representation, has led to several research findings which can be utilized to mitigate the effects of the monoscopic video plus depth map representations. The disadvantages of using this representation and the solutions come across are listed as follows.

- The quality of the rendered stereoscopic views depends on the accuracy of the per-pixel depth values of the original imagery. Therefore, the effects of compression and transmission of depth maps (e.g. introduced artefacts) on the perceived quality need to be thoroughly investigated.
- The visible objects for the rendered virtual left and right views may occlude from the original view. This phenomenon is also known as exposure and disocclusion in Computer Graphics (CG) [21]. This effect can be minimized using Layered Depth Images (LDI) where more than one pair of colour plus depth sequences is transmitted depending on the requirements of the expected quality [26]. However, this approach demands more storage and bandwidth to be used in communication applications. In addition, different hole-filling algorithms (e.g. linear interpolation of foreground and background colour, background colour extrapolation, mirroring of background colour information) can be utilized to recover the occluded areas of the original image sequence [24]. Moreover, the pre-processing/smoothing of depth maps (e.g. use of a Gaussian filter) will avoid this occlusion problem. However, this approach will lead to some geographical distortions of the rendered 3D video scene.
- Certain atmospheric effects (e.g. fog, smoke) and semi-transparent objects are difficult to handle with this approach at the moment.
- The processing overload (e.g. memory, processing power, storage requirements) at the receiver side is high compared to the reconstructing 2D video stream.

The monoscopic video plus depth map representation is widely utilized in research and standardization activities due to its simplicity and adaptability [27] [28] [29]. The ATTEST (Advanced Three-Dimensional Television System Technologies) project consortium is working on 3D-TV broadcast technologies using colour-depth sequences as the main source of 3D video [27]. Recently, ISO/IEC 23002-3 (MPEG-C part 3) finalized the standardization of video plus depth image representations/solutions in order to provide: interoperability of the content, flexibility regarding transport and compression techniques, display independence and ease of integration [28]. Moreover, JVT has identified multi-view video plus depth representation would be a potential candidate for free-view point applications [29]. Due to this wide usage in research and standardization activities, the research carried out in this book utilize the colour plus depth map 3D video representation. This selection would be also supported by the range of advantages associated with this scene representation. For example, the transmission of colour plus depth map would require fewer system resources (bitrate, storage) than the resource requirements for sending left and right views.

Four colour and depth map based stereoscopic video sequences namely Orbi, Interview, Break dance and Ballet are used in the experiments presented in this book. Figure 2.11 shows frames from the original scenes of these test sequences. The Orbi and Interview test video sequences are obtained using a depth/range camera (i.e. Z_{cam}^{TM} camera) are used in the experiments [30]. Orbi is a very complex sequence with camera motion and multiple objects, whereas Interview is a sequence captured with a static camera and featuring a stationary background. The resolution of these two sequences is 720×576 pixels which is the resolution of Standard Definition (SD) TV and original frame rate is 25 frames/s. The rest of the sequences (i.e. Break dance and Ballet) are obtained from the multi-view image sequences generated by the Interactive Visual Media group at Microsoft Research [31]. The fourth camera view and the corresponding depth map computed from stereo are utilized in this experiment [31]. Break dance sequence contains a highly dynamic break dancer in the foreground and a number of supporters with limited motion in the background. In contrast to the Break dance test sequence, Ballet occupies a stationary observer in the foreground and a Ballet dancer operating behind the foreground observer. Both sequences are captured using a stationary camera. The resolution and original frame rate of these two sequences are 1024×768 and 15 frames/s respectively. Due to the use of different colour and depth map sequences (e.g. resolution, frame rate) the results of the experiments will be applicable across most of the application scenarios. Moreover, the findings will be common for all colour plus depth map video representations regardless of the way the material are captured. The issues associated with compression, transmission, display and quality evaluations of this stereoscopic representation are discussed in the following sections.

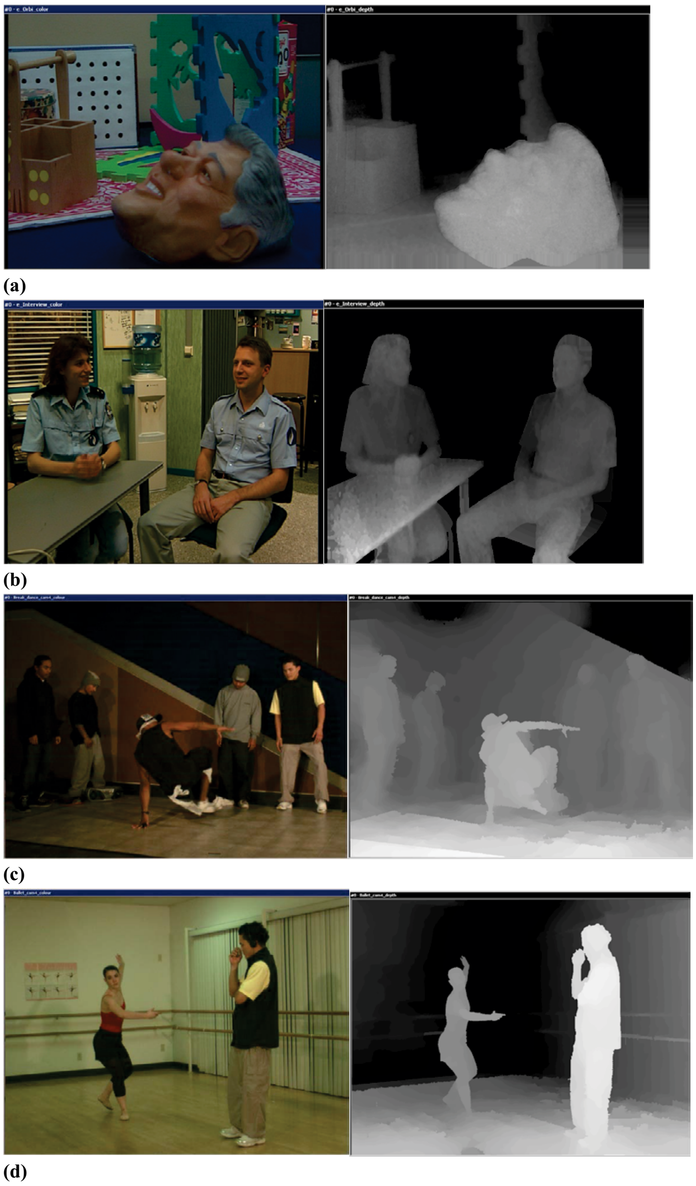


Figure 2.11: Original colour and depth based 3D test sequences; (a) Orbi, (b) Interview, (c) Break-dance, and (d) Ballet

3 Stereoscopic 3D video compression

Your goals for this "Introduction" chapter are to learn about:

- 2D video encoding.
- Scalable video encoding.
- Stereoscopic video encoding.
- Performance evaluation of different encoding approaches for 3D video.

This chapter describes the state of the art video coding approaches for 3D video. An introduction to the 2D video coding algorithms is provided. Moreover, scalable video coding approaches which can be utilized in scaling 2D video applications into immersive video are discussed. Then the potential coding approaches for 3D video in general and more specifically for stereoscopic video are discussed.

3.1 2D Video Coding

The primary aim of video coding is the removal of spatial and temporal redundancies present in raw images captured from a video camera. Video coding allows video to be used in communication applications with reduced storage and bitrate requirements. The block-based transform coding and subband-based decomposition of images are commonly utilized as the basic coding principles. Video coding has been standardized (H.261 in 1990, MPEG-1 Video in 1993, MPEG-2 Video in 1994, H.263 in 1997, MPEG-4 Visual or part 2 in 1998, H.264/AVC in 2004, HEVC in 2013), in order to facilitate the interoperability among different products and applications. The technology advances result in higher compression efficiency, different application support (video telephony-H.261, consumer video on CD-MPEG-1, broadcast of Standard Definition: SD/High Definition: HD TV- MPEG-2 and 4K/8K TV: HEVC) and network compliance (switched networks such as PSTN- H.263/MPEG-4 or ISDN- H.261 and Internet or mobile networks H.263/H.264/MPEG-4). Most of the video coding standards are based on hybrid video coding which employs block matching (i.e. Block Matching Algorithm: BMA) motion compensation and the Discrete Cosine Transform (DCT). The reasons for adopting hybrid video coding approach are that:

- A significant proportion of the motion trajectories found in natural video can be approximately described with a rigid translational motion model.
- Fewer bits are required to describe simple translational motion.
- Implementation is relatively straightforward and amenable to hardware solutions.

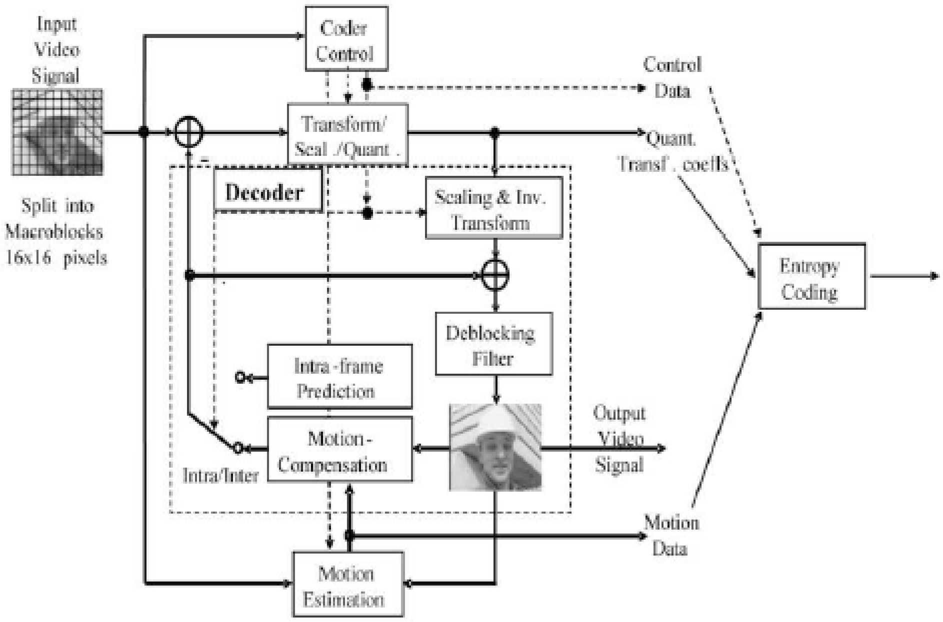


Figure 3.1: Basic structure of a hybrid coder [33]

H.265/High Efficiency Video Coding (HEVC) is the latest video coding technology standardized by the ISO/IEC Moving Picture Experts Group and the ITU-T Video Coding Experts Group [114]. This is primarily based on principles of previous H.264/AVC video coding standard [32]. However, HEVC provides much improved compression efficiency compared to H.264/AVC with added functionality [32]. Figure 3.1 shows the basic structure of a H.264/AVC coder. Similar to the most of the hybrid-video coders, this structure eliminates temporal and spatial redundancies through motion compensation and DCT based transform coding approaches respectively. The high compression efficiency and the network friendliness for interactive and non-interactive video applications are the main achievements in this latest standard [33][34]. The some of the coding features which assist H.264/AVC to gain superior video quality are variable block-size motion compensation with small block sizes, quarter-sample-accurate motion compensation, multiple reference picture motion compensation and in-the-loop de-blocking filter. H.264/AVC consists of two conceptual layers called Video Coding Layer (VCL) and Network Adaptation Layer (NAL). NAL renders a network adaptive bit-stream using the coded bit-stream available at the VCL interface (see Figure 3.2). This close integration of two layers allows H.264/AVC to be used in low bitrate video communication applications across heterogeneous networks.

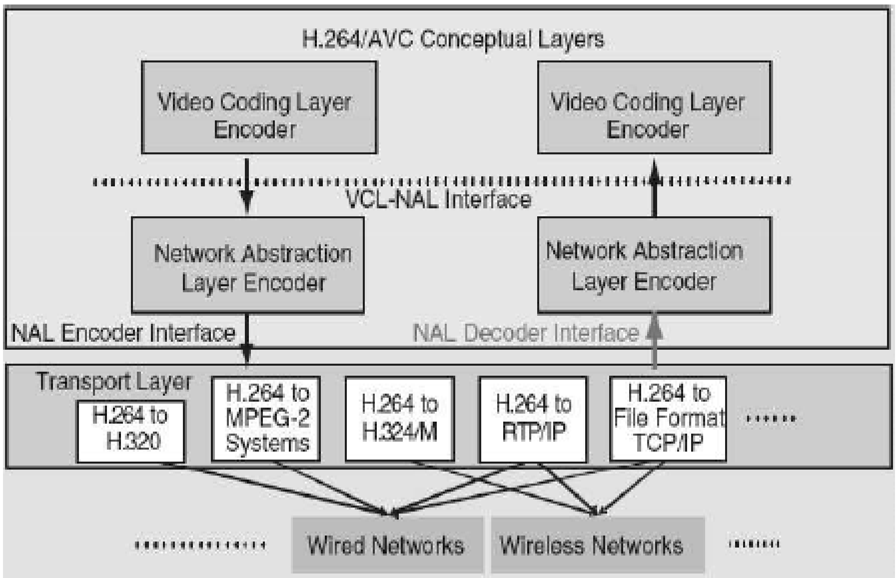


Figure 3.2: H.264/AVC in a transport environment [35]

In addition to the new features used for high compression gain, it consists of several error resilience and concealment features in order to provide more robust and error free video over communication channels. For example it supports slice structure (flexible slice sizes, redundant slices, Flexible Macro-block Ordering-FMO and Arbitrary Slice Ordering ASO), data partitioning, Parameter set structure, NAL unit syntax structure and SP/SI synchronization pictures [33] to be used in error prone environments. The potential tools can be employed in wireless video communication applications and H.264/AVC coded video over best-effort IP networks as described in [35] and [36] respectively.

3.2 Scalable Video Coding

Modern video transmission and storage systems are typically characterised by a wide range of access network technologies and end-user terminals. Varying numbers of users, each with their own time varying data throughput requirements, adaptively share network resources resulting in dynamic connection qualities. Users possess a variety of devices with different capabilities, ranging from cell phones with small screens and restricted processing power, to high-end PCs with high-definition displays. Examples of applications include virtual collaboration system scenarios, as shown in Figure 3.3, where a large, high powered terminal acts as the main control/commanding point and serve a group of co-located users. The large terminal may be the headquarters of the organization and consists of communication terminals, shared desk spaces, displays and various user interaction devices to collaborate with remotely located partners. The remotely located users with a small, fixed terminal will act as the local contact and provide the local information. Mobile units (distribution, surveying, marketing, patrolling, etc.) of the organization may use mobile terminals, such as mobile phones and PDAs, to collaborate with the headquarters.

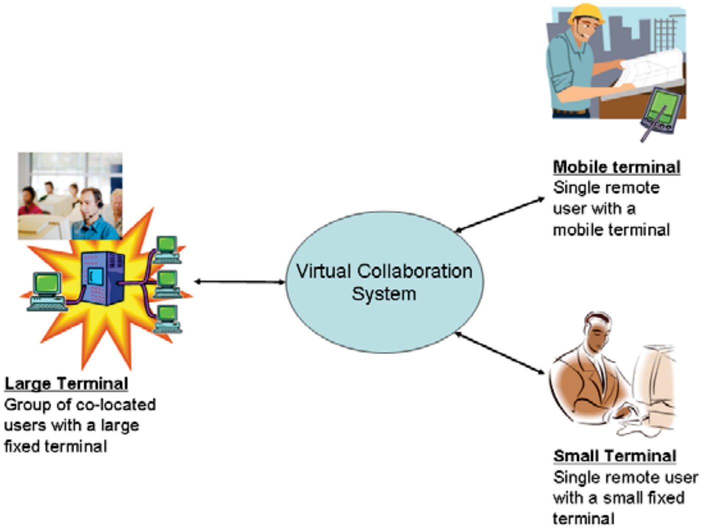


Figure 3.3: Virtual collaboration system diagram

In order to cope with the heterogeneity of networks/terminals and diverse user preferences, the current video applications need to consider not only compression efficiency and quality but also the available bandwidth, memory, computational power and display resolutions for different terminals. The transcoding methods and the use of several encoders to generate different resolution (i.e. spatial, temporal or quality) video streams can be used to address the heterogeneity problem. But above mentioned methods impose additional constraints such as unacceptable delays and increase bandwidth requirements due to redundant data streams. Scalable video coding is an attractive solution for the issues posed by the heterogeneity of today's video communications. Scalable coding produces a number of hierarchical descriptions that provide flexibility in terms of adaptation to user requirements and network/device characteristics. The characteristics of the scalable video coding concept can be utilized to scale the existing 2D video applications into stereoscopic video. For example, colour and depth video can be coded into two scalable descriptors and depending on the receiver terminal capabilities, the users could either render stereoscopic video or shift back to conventional 2D video [37]. This book investigates the adaptability of the scalable video coding concept into backward compatible stereoscopic video applications. Therefore, the background related to scalable video coding is provided.

3.2.1 Scalable coding techniques

At present video production and streaming is ubiquitous as more and more devices are able to produce and distribute video sequences. This brings the increasingly compelling requirement of sending an encoded representation of a sequence that is adapted to the user, device and network characteristics in such a way that coding is performed only once while decoding may take place several times at a different resolution, frame rate and quality. Scalable video coding allows decoding of appropriate subsets of bitstream to generate complete pictures of size and quality dependent on the proportion of the total bitstream decoded. A number of existing video compression standards support scalable coding, such as MPEG-2 Video and MPEG-4 Visual. Due to reduced compression efficiency, increased decoder complexity and the characteristics of traditional transmission systems the above scalable profiles are rarely used in practical implementations. Recent approaches for scalable video coding are based on motion compensated 3D wavelet transform and motion-compensated temporal differential pulse code modulation (DPCM) together with spatial de-correlating transformations [38-41].

The wavelet transform proved to be a successful tool in the area of scalable video coding since it enables to decompose a video sequence into several spatio-temporal subbands. Usually the wavelet analysis is applied both in the temporal and spatial dimensions, hence the term 3D wavelet. The decoder might receive a subset of these subbands and reconstruct the sequence at a reduced spatio-temporal resolution at any quality. The open-loop structure of this scheme solves the drift problems typical of the DPCM-based schemes whenever there is a mismatch between the encoder and the decoder. The scalable video coding based on 3D wavelet transform is addressed in recent research activities [38] [39]. The scalable video coding profiles of existing video coding standards are based on DCT methods. Unfortunately, due to the closed loop, these coding schemes have to address the problem of drift that arises whenever encoder and decoder work on different versions of the reconstructed sequence. This typically leads to the loss of coding efficiency when compared with non-scalable single layer encoding.

In 2007, the Joint Video Team (JVT) of the ITU-T VCEG and the ISO/IEC MPEG standardized a Scalable Video Coding (SVC) extension of the H.264/AVC standard [40]. This SVC standard is capable of providing temporal, spatial, and quality scalability with base layer compatibility with H.264/AVC. Furthermore, this contains an improved DPCM prediction structure which allows greater control over the drift effect associated with closed loop video coding approaches [41].

Bit-streams with temporal scalability can be provided by using hierarchical prediction structures. In these structures, key pictures are coded at regular intervals by using only previous key pictures as references. The pictures between the key pictures are the hierarchical B pictures which are bi-directionally predicted from the key pictures. The base layer contains a sequence of the key pictures at the coarsest supported temporal resolution; while the enhancement layers consist of the hierarchically coded B pictures (see Figure 3.4). A low-delay coding structure is also possible by restricting the prediction of the enhancement layer pictures from only previous frame.

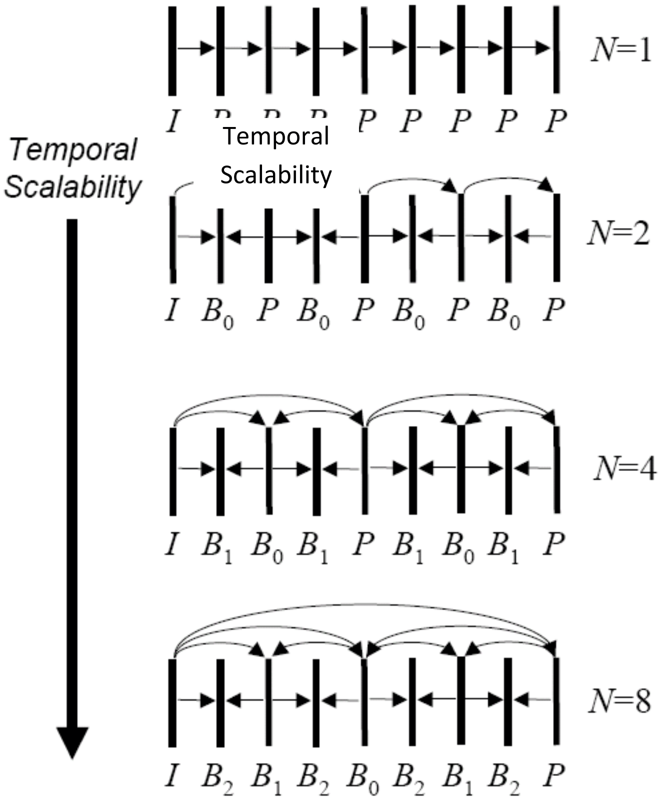


Figure 3.4: Prediction structure for temporal scalability.

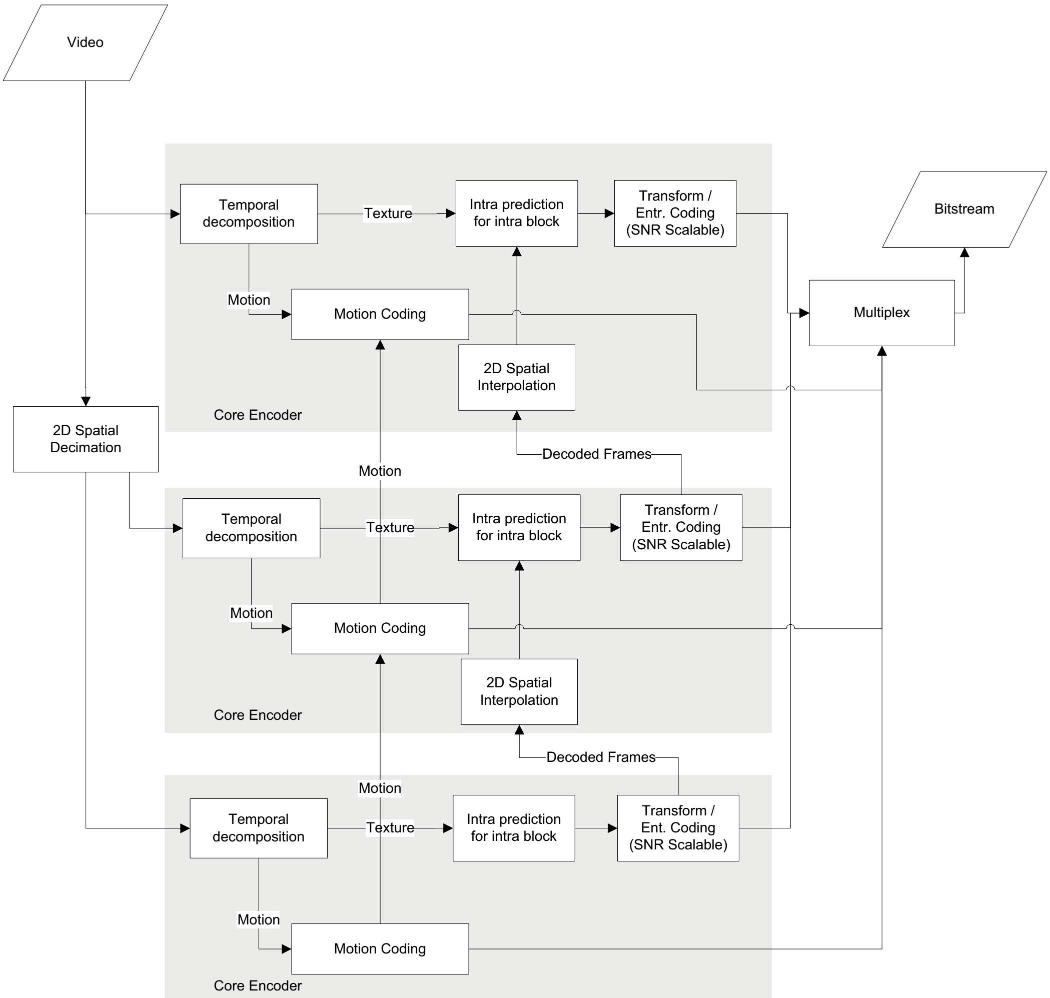


Figure 3.5: Scalable encoder using a multi-scale pyramid with 3 levels of spatial scalability [40]

Spatial scalability is achieved using a multi-layer coding approach in prior coding standards, including MPEG-2 and H.263. Figure 3.5 shows a block diagram of a spatially scalable encoder. In the scalable extension of H.264/AVC, the spatial scalability is achieved with an over-sampled pyramid approach. Each spatial layer of a picture is independently coded using motion-compensated prediction. Inter-layer motion, residual or intra prediction mechanisms can be used to improve the coding efficiency of the enhancement layers. In inter-layer motion prediction, for example, the up-scaled base layer motion data is employed for the spatial enhancement layer coding.

Quality scalability can be considered as a subset of spatial scalability where two or more layers are having similar spatial resolutions but different quality levels. The scalable extension of H.264/AVC also supports quality scalability using coarse-grain scalability (CGS) and medium-grain scalability (MGS). CGS is achieved using spatial scalability concepts with the exclusion of the corresponding up-sampling operations in the inter-layer prediction mechanisms. MGS is introduced to improve the flexibility of bit-stream adaptation and error robustness.

The SVC technology is fast moving towards the deployment of practical, real-time scalable applications [42]. The associated technologies which support scalable applications over communication channels are emerging. For example, Wenger *et al* draft a RTP (Real-time Transport Protocol) payload format for SVC video [43]. This book investigates the possibility of utilizing SVC concept to deliver backward compatible stereoscopic video applications over the networks. Moreover, the scalability features will be employed to encode colour plus depth map stereoscopic video more efficiently with minimal effect for the perceived quality of 3D video.

3.3 3D Video Coding

Scenery captured from different angles/viewpoints simultaneously results a large amount of raw video data. For instance, stereoscopic video, which is one of the simplest forms of 3D video, requires twice the size of storage capacity and double the bandwidth compared to the requirements of 2D video. Thus, 3D video coding is crucial if immersive video applications are to be available for the mass consumer market in the near future. The MPEG Ad-hoc group which worked on exploration of 3DAV (3D Audio-Visual) application scenarios and technologies, identified several MPEG coding tools which are directly applicable for different representations of 3D video and also identified the areas where further standardization is required [5]. The coding approaches for 3D video may be diverse depending on the representation of 3D video. For example, available principles of classical video coding can be utilized to compress pixel-type data including stereo video, multi-view video, and associated depth or disparity maps. However, efficient coding approaches and standardization for some 3D representations such as multi-view video are yet to be discovered. The compression algorithms for 3D mesh models have also reached a high level of maturity. Even though there are efficient compression algorithms for static geometry models, compression of dynamic 3D geometry is still an active field of research. A survey of coding algorithms for different 3D application scenarios is presented in [44]. The coding of image-based 3D representations (e.g. stereoscopic video) is addressed in this book. Henceforth, the encoding of image based 3D representations is discussed in this chapter.

The 3D image/video encoding approaches aim at exploiting inter-view statistical dependencies in addition to the conventional encoding approach for 2D video, which removes the redundancies in the temporal and spatial domains. The prediction of views utilizing the neighbouring views and the images from the same image sequence are shown in Figure 3.6. In general temporal prediction tends to be more efficient compared to inter-view prediction and combined prediction [45]. However, the efficiencies of these prediction methods are varied depending on the frame rate, camera distances and the complexity of content (e.g. motion and spatial details). The JVT (Joint Video Team) standardized the multi-view extension of H.264/AVC [46]. The adopted approach is based on the hierarchical B-frames syntax [see Figure 3.7]. T0, T1,…T100 in Figure 3.7 represent consecutive time instances of image capture whereas S0, S1,…, and S7 represent consecutive camera positions. Each image of the multi-view image sequence is encoded using spatial, temporal or inter-view predictions. This multi-view coding strategy has shown improved coding gain through exploiting inter-view statistical redundancies compared to encoding each view separately. However, the gain is significant only for dense camera settings but not for dome-type (i.e. cameras are set far apart) arrangements. Moreover, the proposed improvements for existing multi-view coding algorithms can be found in the research literature, which enhance the application requirements including random access, low delay and memory optimization [47]. Specific coding tools for MVC such as illumination compensation, view interpolation prediction, inter-view direct mode are also under JVT investigation at present.

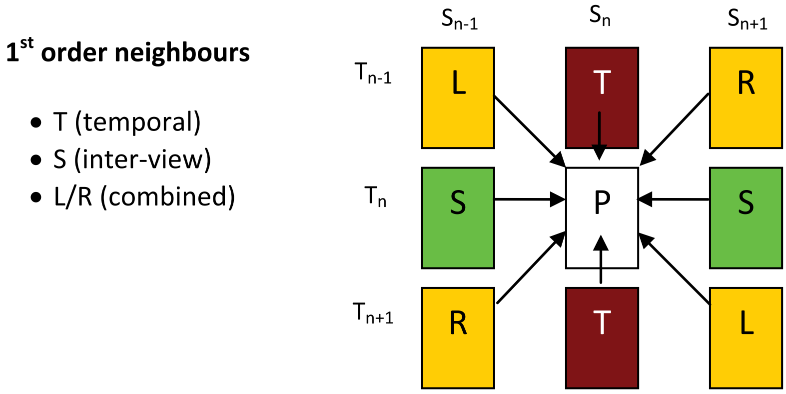


Figure 3.6: Coding of multi-view video

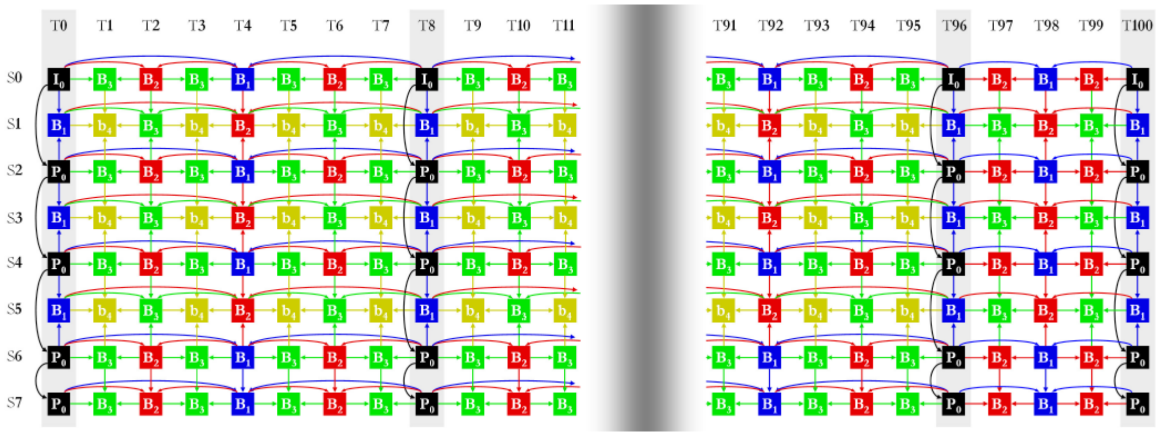


Figure 3.7: Multi-view coding [46]

3.4 Stereoscopic Video Coding

Conventionally, stereoscopic video is captured by two cameras a small distance apart producing two views which differ somewhat of the same object. Therefore, stereoscopic video coding aims to exploit the redundancies present in two adjacent views while removing the temporal and spatial redundancies. For example, an extended H.264 video codec for stereoscopic video, exploiting spatial, temporal, disparity and world-line correlation is described in [48]. Instead of coding the left and right views separately, the coding of left view and the disparity compensated right image has been an efficient method of coding. Disparity estimation algorithms are utilized to identify geometric correspondence of objects in stereo pair [49]. A disparity compensated residual image coding scheme for 3D multimedia applications is described in [50]. The proposed stereoscopic video coding algorithm in [51] employs an object-based approach in order to overcome the artefacts caused by block-based disparity estimation approaches. However, the disparity-compensated prediction approach can not be applicable to colour plus depth video as colour image and associated depth image have different texture information and number of objects. In addition, mixed-resolution video coding for the left and right views is based on the binocular suppression theorem which is based on the response of the human visual system for depth perception. This theory suggests that 3D perception is not affected if one of the view can be in high quality compared to the other view [134]. The high quality view drives 3D perception by compensating the slight loss in the other view. H.264/AVC based stereoscopic video codec described in [52] utilizes asymmetric coding of the left and right images in order to improve the compression efficiency. The results in [52] show that stereoscopic video can be coded at a rate about 1.2 times the monoscopic video using spatial and temporal filtering of one image sequence. Moreover, a 3D video system based on H.264/AVC view coding described in [53] examines the bounds of asymmetric stereo video compression. The mixed-resolution coding concept can be applied to colour plus depth video in different perspectives, i.e. to code the colour and depth video based on their influence towards better perceptual quality. For example, the depth image can be coded differently compared to the corresponding colour image [130].

The multi-view profile (MVP) of the MPEG-2 video coding standard can be utilized in coding stereoscopic video with Disparity-Compensated Prediction (DCP). This profile is defined based on the Temporal Scalability (TS) mode of MPEG-2 [54]. However, due to the coarser disparity vectors which are defined on a block-by-block basis of 16x16 pixels, the inter-view prediction error is larger compared to the motion compensated error [55]. Therefore, the gain obtained with this coding approach is not significant compared to the coding of left and right views separately. However, the proposed improvements (e.g. global displacement compensated prediction [56]) for MPEG-2 multi-view profile have shown improved picture quality at low bitrates. Due to the disparity estimated prediction structure of the MPEG-2 multi-view profile this approach is not suitable for encoding colour plus depth video.

The MPEG-4 Multiple Auxiliary Component (MAC) is a generalization of the grayscale shape coding [57]. This allows image sequences to be coded with a Video Object Plane (VOP) on a pixel-by-pixel basis which contains data related to video objects such as disparity, depth and additional texture. However, this approach needs to send the grayscale/alpha shape with any other auxiliary component (e.g. disparity) and as a result the coding efficiency is affected. Even though MPEG-4 MAC is suitable to encode colour plus depth video due to the presence of grayscale reduces the coding gain. Therefore, the performance of a modified MPEG-4 MAC coder with no alpha plane is analyzed in this book in the last part of this chapter.

The multi-view codec developed by JVT can also be utilized in stereoscopic video coding with inter-view disparity compensation [46]. However, this method is not applicable for colour plus depth stereoscopic video due to the availability of different texture information in each image sequence. Consequently, the amount of redundancies which can be exploited through inter-layer prediction is minimal. The recent interest on multi-view plus depth representation may suitable to encode monoscopic video plus depth information [29]. However, these approaches are developed for interactive multi-view applications such as free-viewpoint video and may not so effective towards more simple representations like stereoscopic video.

The MPEG-4 Animation Framework eXtension (AFX) standard ISO/IEC 14496-16 defines two image-based depth representations that can be used to encode colour plus depth video [58]. The two categories are the simple Depth Image (DI) and Layered Depth Image (LDI) representation which is utilized to store occluded areas of the original image. However, this approach seems very much computer graphics oriented and suitable for synthetic image synthesis [58]. Therefore, this approach is not suitable for encoding natural scenes using monoscopic video plus depth representations of 3D video.

The quality of the virtual left and right views generated through DIBR is dependent on the quality of the colour video, associated depth map and the camera geometry. The relationship can be attributed to the Quantization Parameters (QPs) of each image and the virtual camera position (see Equation 3.1).

$$\text{Quality} = f\,(QP_{colour}, QP_{depth}, \text{Camera position}) \qquad \textbf{Equation 3.1}$$

However, the effect of colour image quality is more influential with this approach than the depth image quality as texture information is directly viewed by the users. Therefore, effects of colour and depth coding on the perceptual quality needs to be studied thoroughly. This chapter investigates the perceptual bounds of depth map coding for stereoscopic video applications. For example, spatially down-sampled depth maps would be accurate enough to generate good quality left and right views. Moreover, the efficient coding architectures for colour plus depth map coding and their suitability for communication applications are reported in this chapter.

Due to the characteristics of the depth image sequence (e.g. smoothness of real world objects), it can be efficiently compressed with the existing coding algorithms than the coding of corresponding monoscopic video sequence [59]. According to [59], H.264/AVC outperforms MPEG-2 and MPEG-4 video coding standards in depth map coding. For example, depth maps can be encoded with H.264/AVC at a bitrate less than 20% of the MPEG-2 coded colour image sequence for Standard Definition (SD) resolution image sequences. The following subsection describes potential encoding approaches for stereoscopic 3D video.

3.4.1 Exploration of Efficient Stereoscopic Video Coding Configurations

Simulcast configurations:

- This approach generates two bit-streams
- Needs multiplexing and de-multiplexing modules after and before the video codec
- Inter-view redundancies are not removed, and hence less efficient

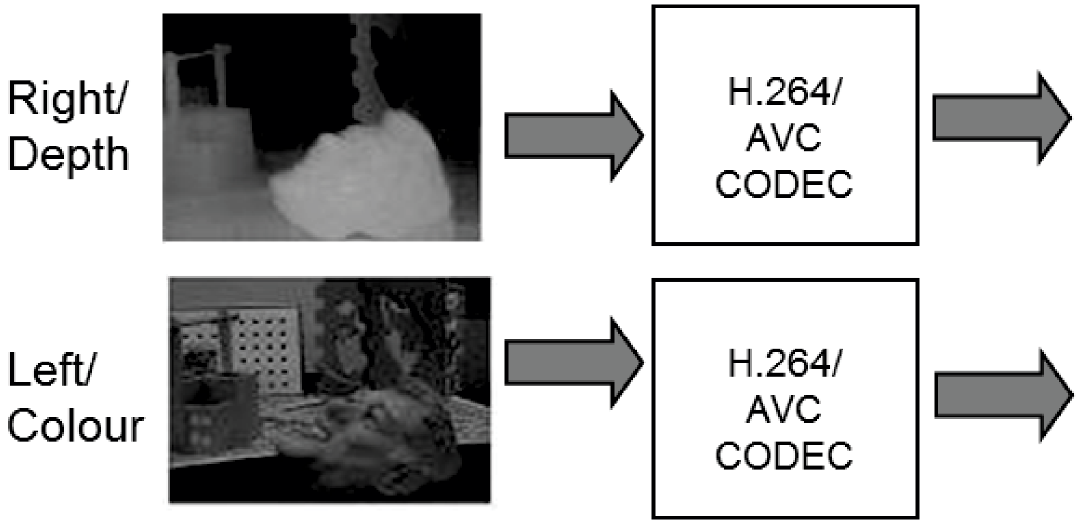


Figure 3.8: Parallel/Simulcast encoding approach for 3D video

Pre-processing configurations: Also known as frame packing 3D format. This refers to the combination of two frames, one for the left eye and the other for the right eye, into a single "packed" frame that consists of these two individual sub-frames. The key difference of a Frame Packing signal is that each sub-frame for each eye is still at full resolution, i.e., 1920×1080 for a 1080p Frame Packing signal, and 1280×720 for 720p Frame Packing 3D content.

Side-by-Side approach:

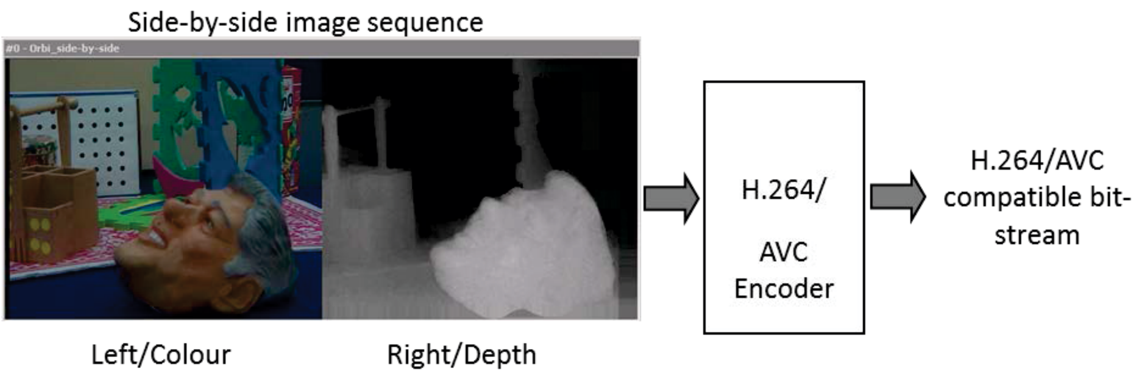


Figure 3.9: Frame packed configuration- side-by-side format

Top-Bottom approach:

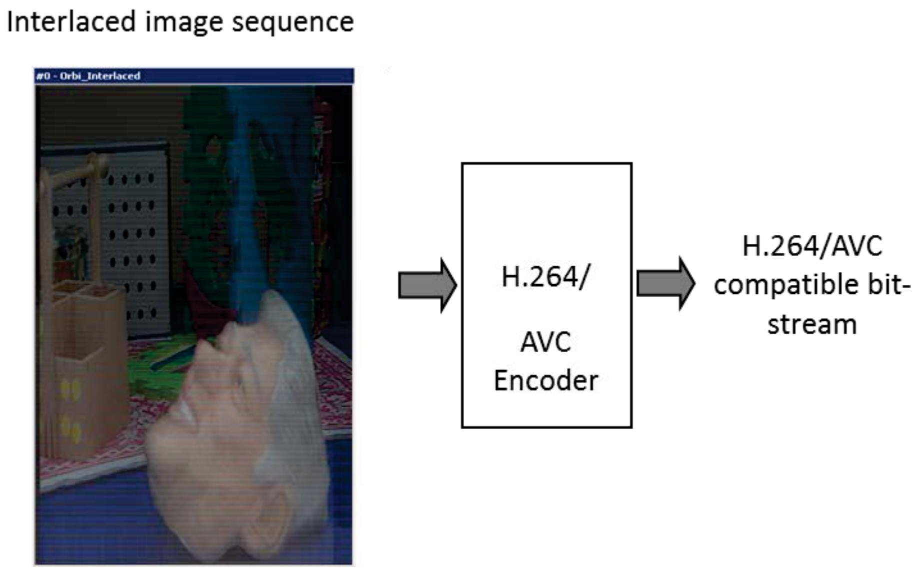


Figure 3.10: Frame packed configuration- top-bottom format Type 1 (interlaced format)

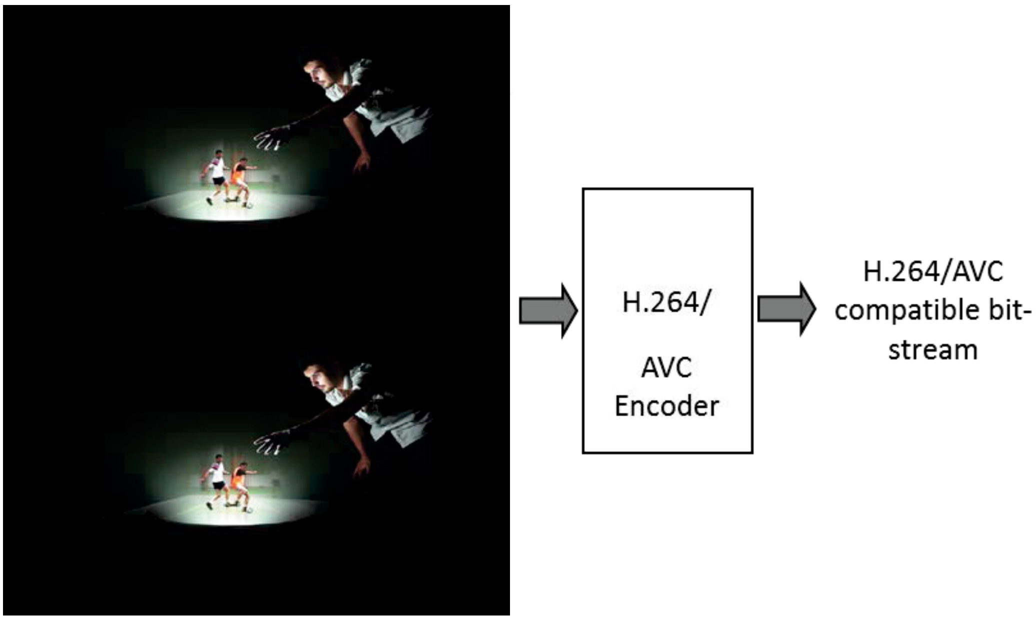


Figure 3.11: Frame packed configuration- top-bottom format Type 2

- This approach generates a single output bit-stream
- A single coded can be utilized for encoding
- Inter-view redundancies are not removed, and hence less efficient

Frame compatible 3D: Both Frame Compatible 3D and Frame Packing 3D formats involve forming a single frame that contains "sub-frames" for the left and right eye. In both cases, the Sub-frames can be packaged together into a single frame via the Side-by-Side 3D format or the Top-and-Bottom 3D format. The key difference of a Frame Compatible signal is that each sub-frame for each eye is down sampled along one axis to lower the resolution of each sub-frame along one axis. As a result, the total dimension of a Frame Compatible Frame is the same as a regular 2D HD frame (since each sub-frame has half the resolution along either the horizontal or vertical dimension). This is the reason this format is called Frame Compatible.

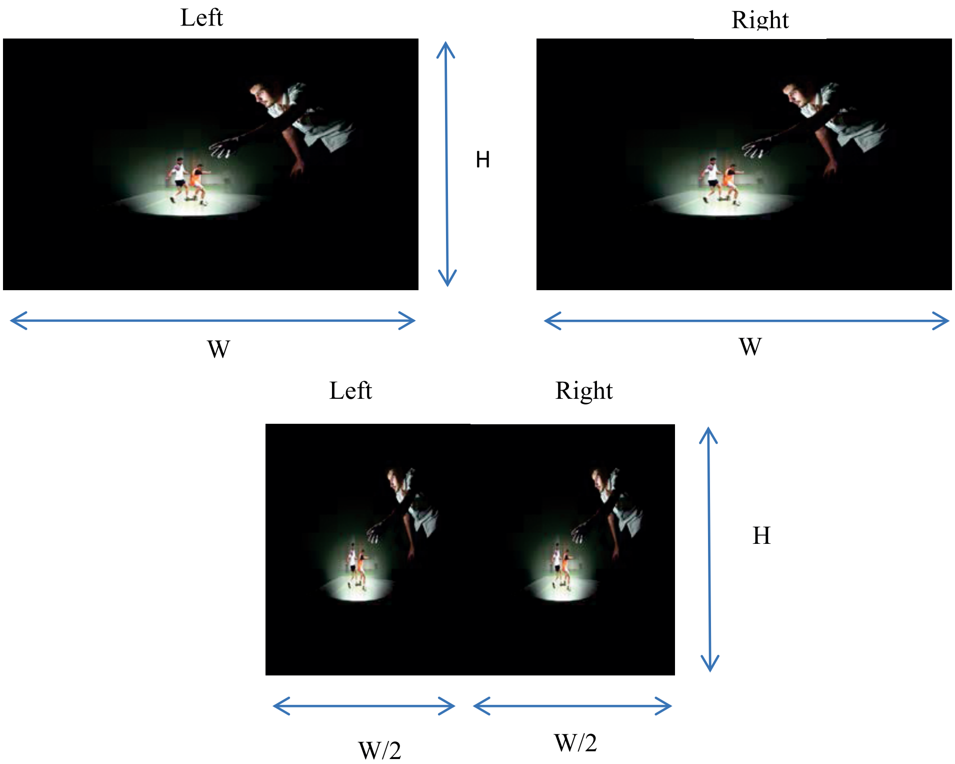


Figure 3.12: Frame compatible format for 3D video

Layered encoding configurations (e.g., MPEG-2 TS, Scalabel H.264/AVC, Multi-View Coding):

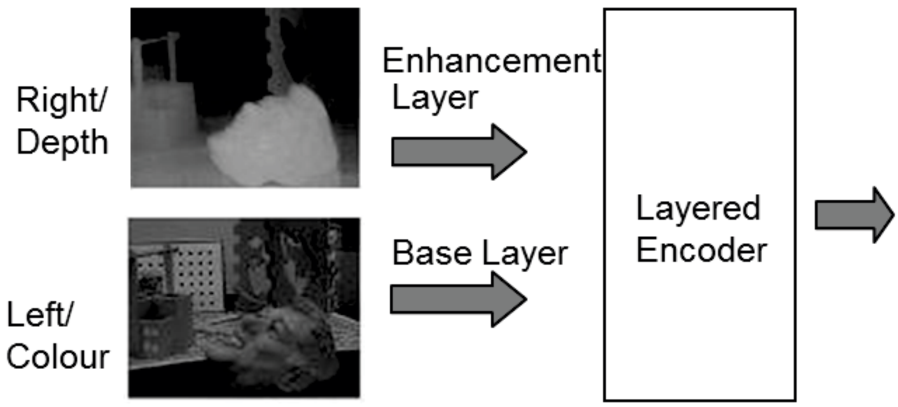


Figure 3.13: Layered encoding approach for 3D video

- Generates a single bit-stream
- Removes inter layer redundancies
- Backward compatibility (i.e., the base layer is compatible for 2D video decoding)
- Asymmetric coding support using temporal, spatial and quality scalability options available with H.264/SVC

Two layered encoding approaches for stereoscopic 3D video are describe in detail below.

Stereoscopic Video Coding Configuration Based on Scalable Video Coding (SVC): The Scalable Video Coding (SVC) encodes video data in different resolutions (spatial/temporal) and quality (refer to the SVC section above). Depending on the receiver capabilities the users can decode the full-resolution image or sub-version of the image. Most of the SVC methods utilize the layered coding approaches to encode scalable image sequences. This study proposes a stereoscopic video coding configuration based on the layered coding architecture of SVC. The proposed method is implemented on the scalable extension of H.264/AVC. Scalable H.264/AVC is developed and standardized by JVT and supports spatial, temporal and quality scalability for video coding [40]. In this coding configuration, the colour and depth image sequences are coded at the base and enhancement layers respectively as shown in Figure 3.14. In addition, this layered coding architecture can also be utilized to encode left and right view based stereoscopic video. The coding of left and right views with this coding configuration is considered in this chapter during the performance comparison of colour plus depth coding vs. left and right view coding.

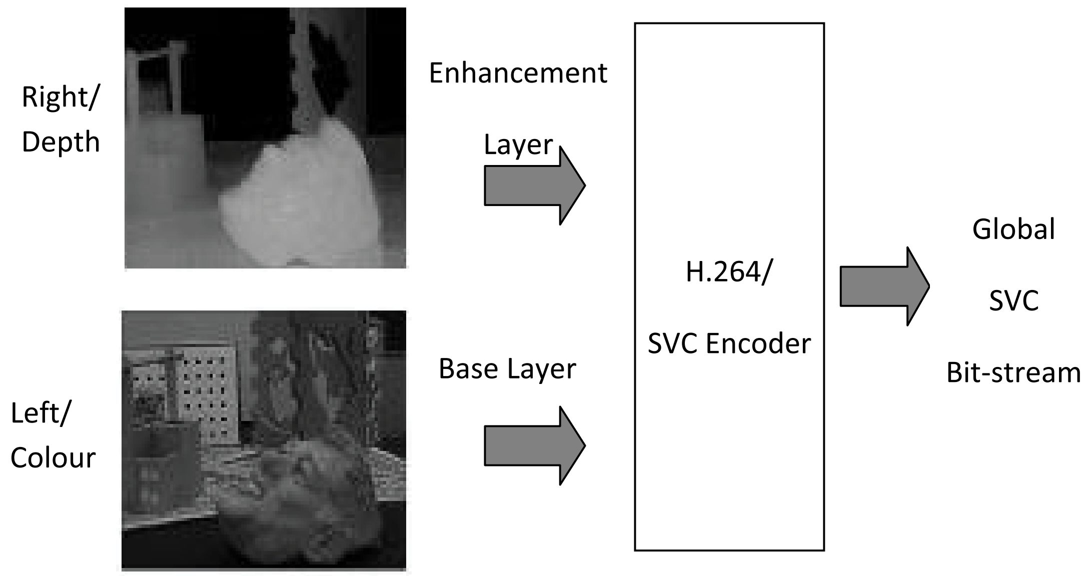


Figure 3.14: Stereo video coding configuration based on scalable H.264/AVC.

As the base layer of this configuration is compatible with H.264/AVC decoders, users with a H.264/AVC decoder will be able to decode the colour image sequence, whereas users with a SVC decoder and 3D rendering hardware will be able to decode the depth map sequence and experience the benefits of stereoscopic video. Therefore, the backward compatibility nature of this scalable coding configuration can be employed to enhance or scale existing 2D video applications into stereoscopic video applications. Furthermore, this configuration can be utilized to exploit asymmetric coding of the colour and depth map image sequence since scalable H.264/AVC supports a range of temporal, spatial and quality scalable layers. For instance, during encoding, the resolution of colour and depth image sequences can be independently changed based on their inherent characteristics, without affecting the perceptual qualities of 3D video. The proposed asymmetric coding methods with this scalable coding configuration are presented in [131]. Furthermore, inter-layer redundancies can be exploited based on the correlation between the colour and depth images. The results presented here make use of the adaptive inter-layer prediction option in the JSVM (Joint Scalable Video Model) reference software codec, which selects the coding mode for each MB (Macro-Block) using Rate-Distortion (R-D) optimization. Moreover, this coding configuration complies with the ISO/IEC 23002-3 (MPEG-C part 3) standard, which specifies the inter–operability among applications based on colour plus depth/disparity sequences [28]. The output of this coding configuration is a single SVC bit-stream which can be partially decoded. Therefore, the transmission and synchronization of this content over communication channel is not difficult compared to sending colour and depth streams separately. Furthermore, the supportive technologies (e.g. RTP format for SVC) are emerging to support the delivery of SVC data streams over communication channels [43]. Subsection 3.5.1 discusses the coding efficiency of the proposed SVC configuration compared to MPEG-4 MAC and H.264/AVC based stereoscopic video coding approaches.

Stereoscopic Video Coding with MPEG-4 Multiple Auxiliary Components (MAC): The MAC (Multiple Auxiliary Components) is added to Version 2 of the MPEG-4 Visual part [57] in order to describe the transparency of video objects. The MAC is defined for a video object plane (VOP) on a pixel-by-pixel basis and contains data related to video objects such as disparity, depth, and additional texture. As MPEG-4 MAC allows the encoding of auxiliary components (e.g. depth, disparity, shape) in addition to the Y, U and V components present in 2D video, this coding approach can be utilized to compress colour plus depth stereoscopic video [108] or disparity compensated stereoscopic video [109]. The coding of monoscopic video plus depth map with MPEG-4 MAC is illustrated in Figure 3.15. Even though the encoding of shape information is compulsory with any other auxiliary data stream (i.e. depth map), the MAC coder utilized in this book is modified to remove the shape coding requirement of MPEG-4 MAC. Consequently, this coding configuration is expected to present improved coding efficiency compared to the original MPEG-4 MAC coder [55].

MPEG-4 MAC is a good mechanism for generating one-stream stereoscopic video coding output. The one-stream approach facilitates end-to-end video communication applications without a system level modification (avoid multiplexing and de-multiplexing stages for different streams) and can be in compliant to the ISO/IEC 23002-3 (MPEG-C part 3) standard. In this work, the R-D performance of MPEG-4 MAC coded colour plus depth stereoscopic video is compared against results obtained with the H.264/AVC and scalable H.264/AVC video coding standards. However, it should be noted that the MPEG-4 MAC utilized is based on the ISO/IEC 14496-2 standard (MPEG-4 Visual), rather than the AVC technology, which is ISO/IEC 14496-10.

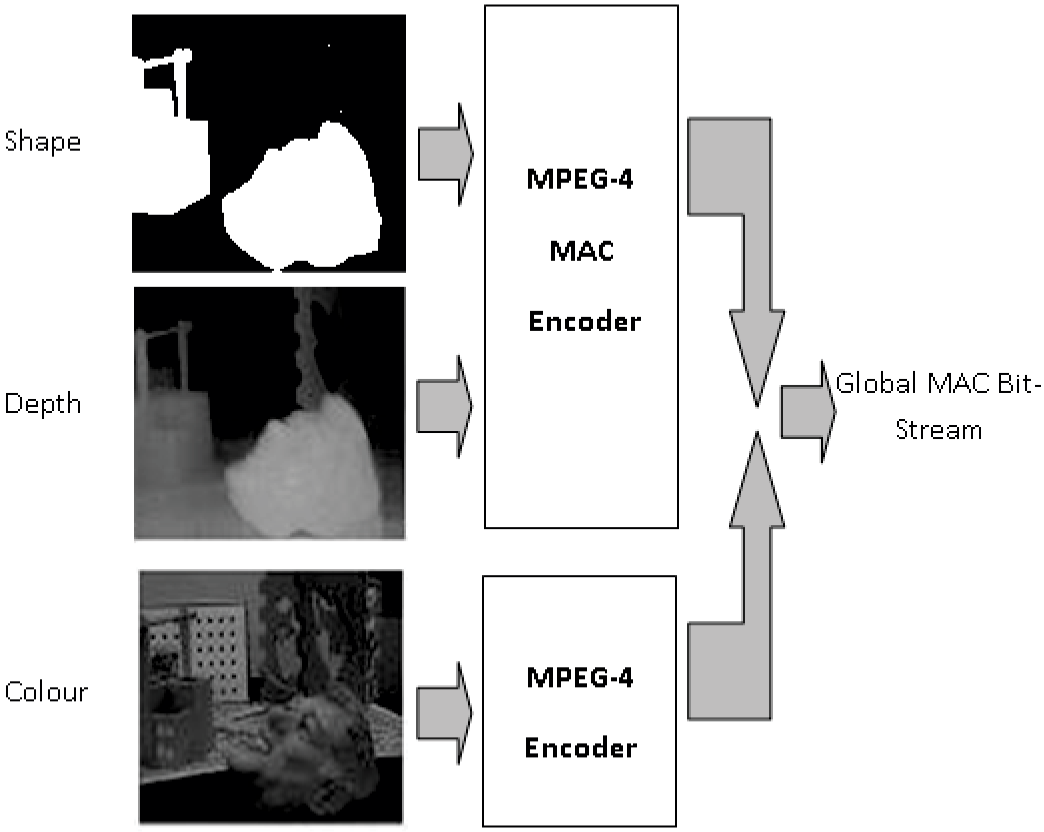


Figure 3.15: Stereo video coding configuration based on MPEG-4 MAC.

3.5 Performance analysis of different encoding approaches for colour plus depth based 3D video and comparison of left and right view encoding vs. colour plus depth map video encoding

The following encoding configurations are used in this evaluation;

- Layered encoding with H.264/AVC Scalable Video Coding
- MPEG-4 MAC Encoding
- H.264/AVC Encoding with side-by-side stereoscopic 3D video (not frame compatible format)

The Orbi and Interview test sequences are utilised to obtain the Rate-Distortion (R-D) results. The experiments are carried out using CIF (Common Intermediate Format, 352×288) format image sequences in order to evaluate the performance of low bitrate 3D video applications. As the original resolution of these sequences is 720×576, the image sequences are initially cropped (i.e. 704×576) and then down-sampled into CIF (352×288) format video. During the down-sampling process, four nearest pixel values of the original format are averaged to represent the pixel value of the down-sampled image.

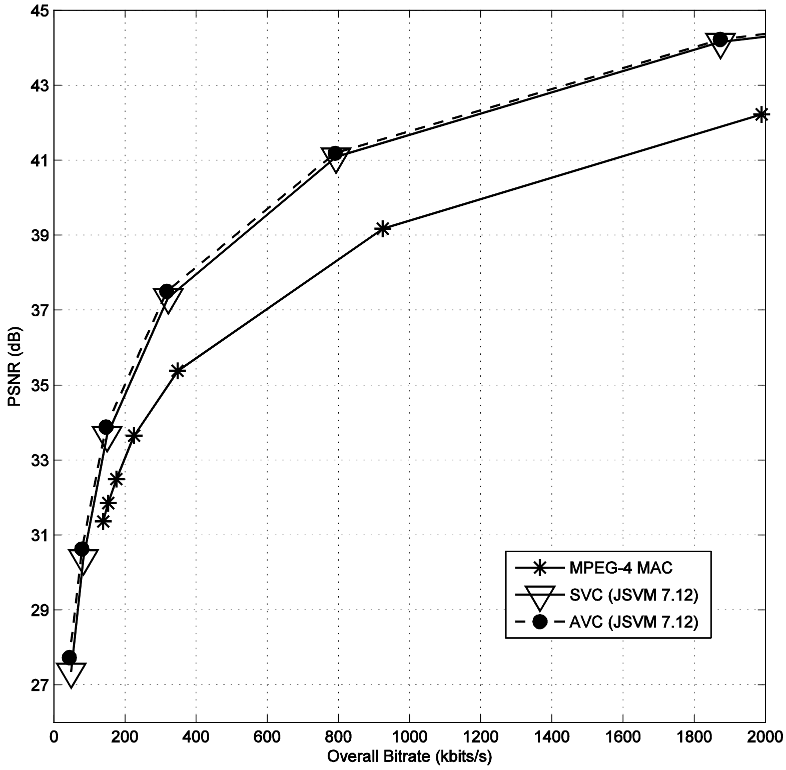
The image sequences are encoded with the three encoding configurations based on the existing video coding standards, MPEG-4 MAC, H.264/AVC (using the scalable H.264/AVC single layer coding) and the scalable extension of H.264/AVC. The base layer encoder of the scalable H.264/AVC (i.e. encode using the single layer configuration) is utilized to obtain H.264/AVC coding results as the base layer bit-stream is backward compatible for H.264/AVC, and to avoid differences arising from the different software implementations of H.264/AVC and the scalable extension of H.264/AVC. The basic encoding parameters used are shown in Table 3.1. The Quantisation Parameter (QP) in the configuration file is varied to obtain the bitrate range shown in the R-D curves. The same QP is used for encoding both base and enhancement layers of scalable H.264/AVC. The R-D curves show the image quality measured in PSNR against the resulting average bitrate. Experiment 1 compares the coding performance of the stereoscopic video coding architectures. The coding performance of colour and depth video vs. left and right video with the proposed SVC coding configuration is discussed in Experiment 2.

Encoding Parameter	Value
Test sequences	Orbi, Interview
No. of frames	300
Sequence format	IPPP…
Reference frames	1
Search range	16 pixels
Entropy coding	VLC (Variable Length Coding)

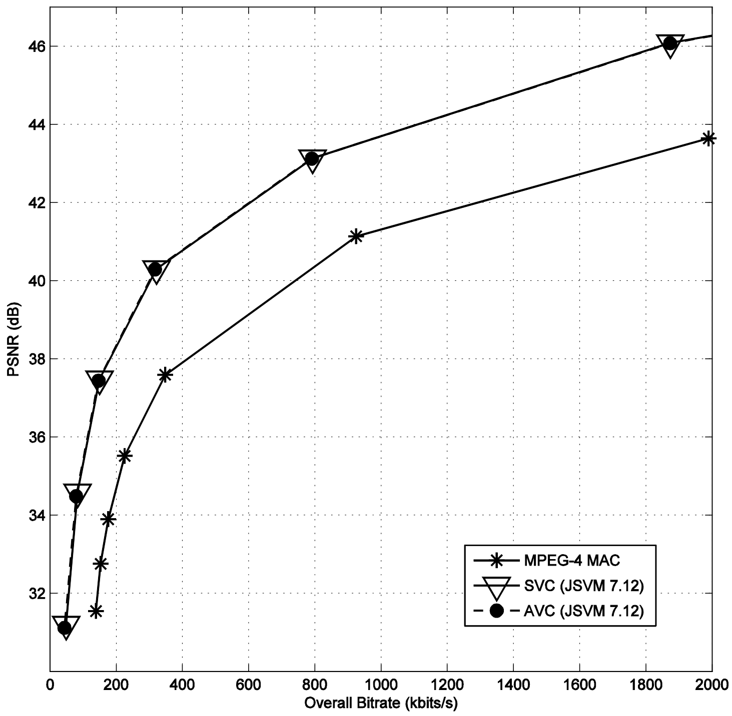
Table 3.2: Encoding parameters

3.5.1 Experiment 1: Comparison of Stereoscopic Video Encoding Configurations

The R-D performance of Orbi and Interview colour and depth sequences using MPEG-4-MAC (layered encoding), H.264/AVC (frame packing, side-by-side format) and scalable H.264/AVC coding (layered encoding) configurations are shown in Figures 3.16 and 3.17 respectively. All of the results are plotted against the overall bitrate (output bitrate of the SVC codec), which includes all of the overhead, texture, colour, motion vector bits of both colour and depth map video. In order to highlight the R-D performance at a range of bitrates, the final bitrate is shown from 0 Kbits/s to 2Mbits/s. The H.264/AVC coded stereoscopic video sequences (i.e. side-by-side images) are separated into colour and depth map video in order to calculate the PSNR with respect to their original colour and depth image sequences.

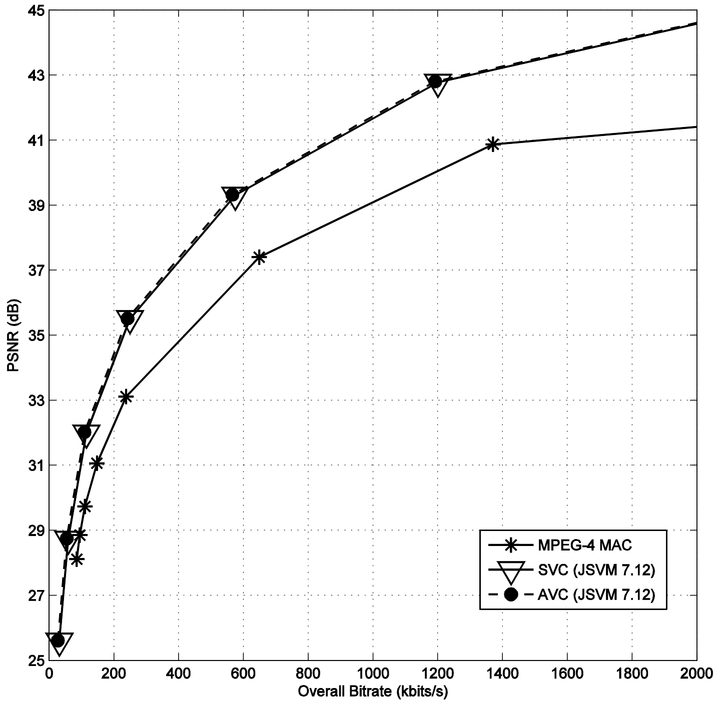


(a)

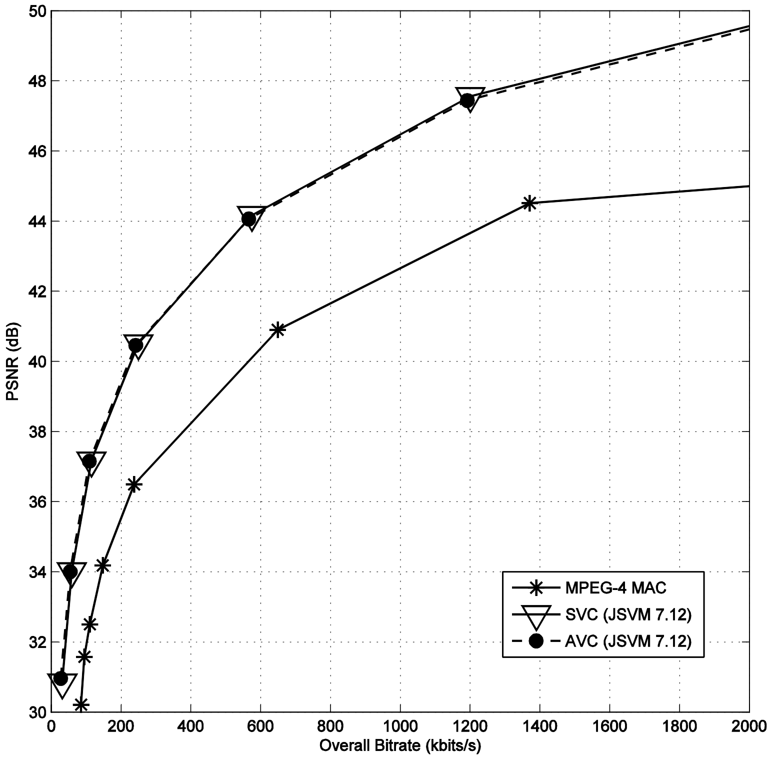


(b)

Figure 3.16: R-D curves for Orbi sequence (a) Colour image sequence (b) Depth image sequence.



(a)



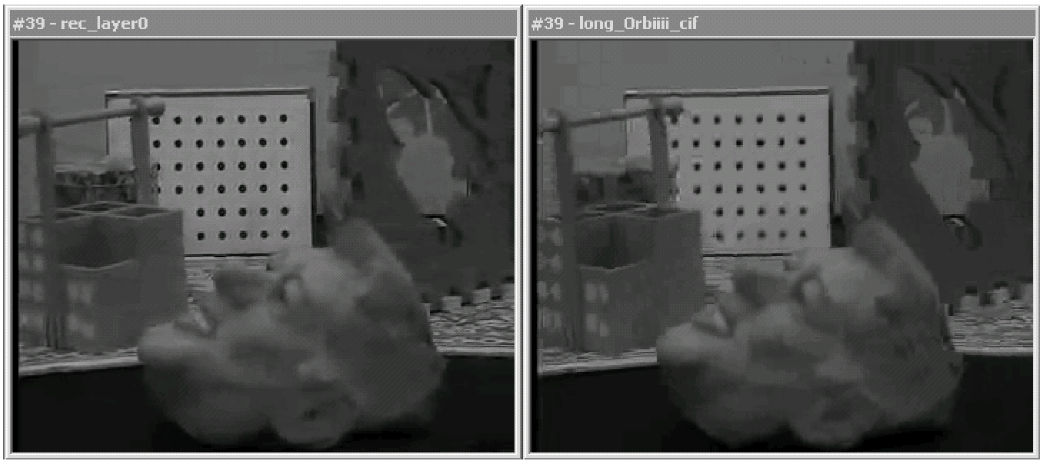
(b)

Figure 3.17: R-D curves for Interview sequence (a) Colour image sequence (b) Depth image sequence.

The R-D curves for both Orbi and Interview sequences show that the proposed configuration with scalable H.264/AVC performs similar to the H.264/AVC configuration, and outperforms the MPEG-4 MAC based configuration at all bitrates. The SVC configuration has not outperformed the H.264/AVC configuration due to the negligible usage of inter-layer prediction between the base and enhancement layers and extended header data associated with SVC. Even though, common object boundaries are present in colour and depth map sequences, the depth image sequence has different texture information, and less number of objects and high frequency components compared to the colour video. Consequently, the number of inter-layer coding modes used is very small during the encoding process. In addition, the R-D performance of the scalable H.264/AVC configuration is slightly affected due to the overhead bits added to the SVC bit-stream to recognize enhancement layer data. However, the flexibility of the SVC configuration in stereoscopic video coding, such as asymmetric coding support (temporal, spatial and quality scalability for depth image sequences), single bit-stream output, and backward compatibility, facilitates end-to-end stereoscopic video communication chain to a greater extent. The flexible macro-block (MB) sizes and skipped MB features available in H.264/AVC standard have helped to achieve better coding performance with the H.264/AVC and scalable H.264/AVC based configurations compared to the MPEG-4 MAC configuration at all bitrates. Furthermore, it can be observed that the configurations based on H.264/AVC provide reasonable image quality at very low overall bitrates compared to the MPEG-4 MAC configuration.

The H.264/AVC and scalable H.264/AVC configurations outperform the MPEG-4 MAC configuration by a considerable margin for depth image quality at all overall bitrates (see Figure 3.16 (b) and 3.17 (b)). The fewer number of objects and unavailability of high frequency components help to achieve superior depth map quality for H.264/AVC based configurations. These smooth depth images can be highly compressed using flexible MB sizes and skipped MB modes available in H.264/AVC [110] compared to other MPEG video coding standards. The gain is more visible in the Interview sequence, which has less motion and stationary background.

The subjective image qualities of the Orbi colour and depth map sequences are illustrated in Figures 3.18 and 3.19 respectively. The image sequences are obtained at an overall bitrate of 150 Kbits/s, utilizing the coding configurations based on scalable H.264/AVC and MPEG-4 MAC. According to Figure 3.18, subjective quality of the SVC coded colour image is better compared to that of the MPEG-4 MAC coded colour image. This is more visible in the depth image sequences as shown in Figure 3.19. The scalable H.264/AVC coded depth image demonstrates a sharp and better image quality compared to that of the MPEG-4 MAC coded depth image sequence at the given low bitrate of 150 Kbits/s



(a) **(b)**

Figure 3.18: Subjective image quality of the Orbi colour sequence at an overall bitrate of 150 Kbits/s (a) Scalable H.264/AVC configuration (b) MPEG-4 MAC configuration.

(a) (b)

Figure 3.19: Subjective image quality of the Orbi depth sequence at an overall bitrate of 150 Kbits/s (a) Scalable H.264/AVC configuration (b) MPEG-4 MAC configuration.

Table 3.3 shows the image quality of both sequences at an overall bitrate of 200 Kbits/s. This shows that the proposed stereoscopic video coding configuration based on scalable H.264/AVC provides superior quality compared to the MPEG-4 MAC configuration at overall bitrates as low as 200Kbits/s. The high performance and flexible features (backward compatibility and temporal, spatial and quality scalability) associated with the SVC architecture can be employed to scale low bitrate conventional video applications into stereoscopic video applications. Furthermore, at an overall bitrate of 200Kbits/s the Orbi depth image sequence can be coded at 49% of the Orbi colour image bitrate using the proposed configuration based on scalable H.264/AVC. The depth image bitrate requirements could be further reduced by using a high QP value or reduced temporal or spatial scalability at the enhancement layer (depth image) encoder without affecting the perceptual quality of the stereoscopic video application. Moreover, the occlusion problems associated with the DIBR method can also be resolved using this proposed SVC configuration with the use of several pairs of colour plus depth images, which is known as Layered Depth Images (LDI).

Encoding configuration	**Orbi Y-PSNR (dB)**		**Interview Y-PSNR (dB)**	
	Colour	**Depth**	**Colour**	**Depth**
Proposed (Scalable H.264/AVC)	34.74	38.31	34.22	39.29
MPEG-4 MAC	33.05	34.68	32.25	35.52
H.264/AVC	35.01	38.33	34.41	39.41

Table 3.3: Image quality at an overall bitrate of 200 Kbits/s

3.5.2 Experiment 2: Colour plus Depth Coding vs. Left and Right View Coding

This experiment compares the R-D performance of colour and depth image sequences vs. left and right image sequences encoding using the efficient scalable H.264/AVC configuration. In order to produce left and right image sequences, the Orbi and Interview sequences are projected into virtual left and right image sequences using the DIBR as described in Equation 3.2 and coded as the base and enhancement layers with the SVC coding configuration. The rendered left and right image sequences can be considered as stereoscopic video captured using a stereo camera pair. Then, the coded colour and depth image sequences at the base and enhancement layers (i.e. using the scalable H.264/AVC) are also converted to virtual left and right video to be compared with the coded left and right video.

$$P_{pix} = -x_B \frac{N_{pix}}{D}\left[\frac{m}{255}(k_{near} + k_{far}) - k_{far}\right] \qquad \textbf{Equation 3.2}$$

Where, N_{pix} and x_B are the number of horizontal pixels of the display and eye separation respectively. The depth value of the image is represented by the N-bit value m. k_{near} and k_{far} specify the range of the depth information behind and in front of the picture, relative to the screen width N_{pix} respectively. The viewing distance is represented by the parameter value D.

Figures 3.20 and 3.21 show the R-D performance of Orbi and Interview sequences respectively at low bitrates up to an overall bitrate of 500 Kbits/s. Both Orbi and Interview sequences demonstrate better performance for colour and depth image coding than coding projected left and right view video with scalable H.264/AVC. Due to the characteristics of depth image sequences, they can be highly compressed using H.264/AVC coding tools [110]. As a result the coded depth content requires less average bitrate compared to its corresponding colour image sequence. The coded colour and depth content can be squeezed into a given target bitrate with less degradation in colour and depth image quality, which is vital to render two high quality left and right sequences using DIBR method. In case of left and right view coding with scalable H.264/AVC, both image sequences require considerable amount of bitrate due to the availability of similar texture information in both images. Furthermore, the amount of disparity between the projected left and right images has hindered the use of adaptive inter-layer prediction more effectively between the base and enhancement layers of the proposed SVC configuration. Consequently, left and right view coding has failed to achieve improved quality at low bitrates. However, at higher bitrates left and right video coding may achieve better performance due to the possibility of squeezing good quality left and right sequences into a given bitrate.

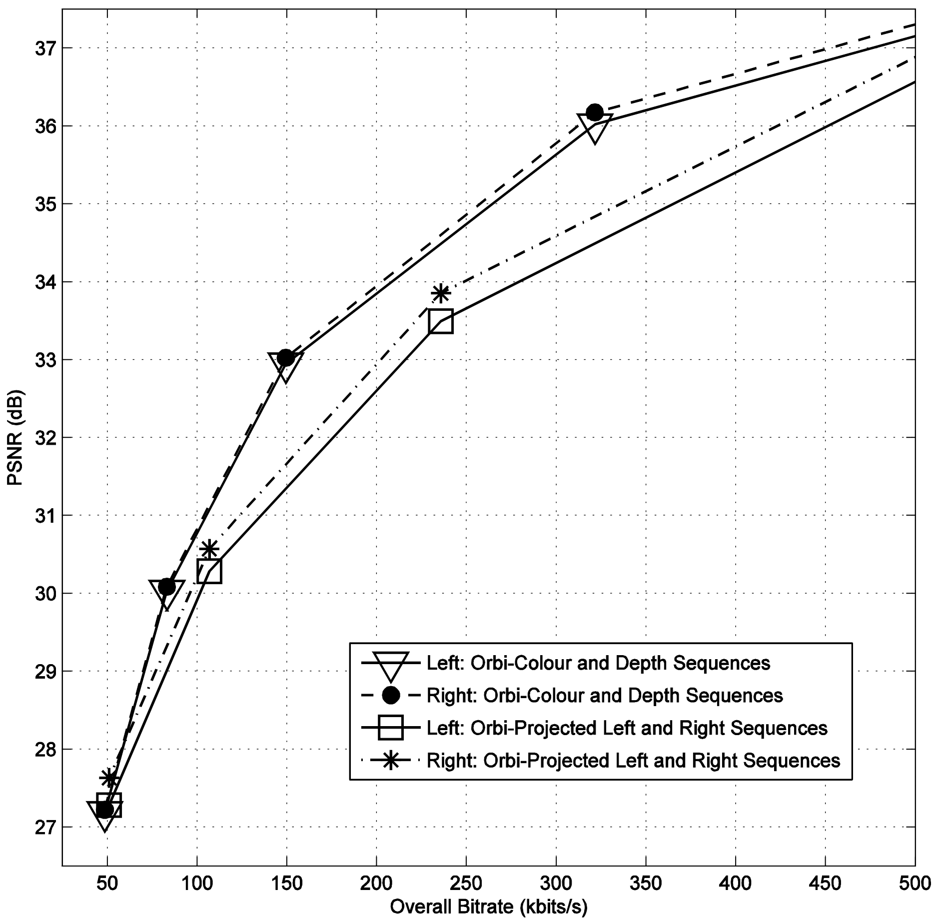


Figure 3.20: R-D curves for Orbi (using colour and depth sequences and projected left and right sequences).

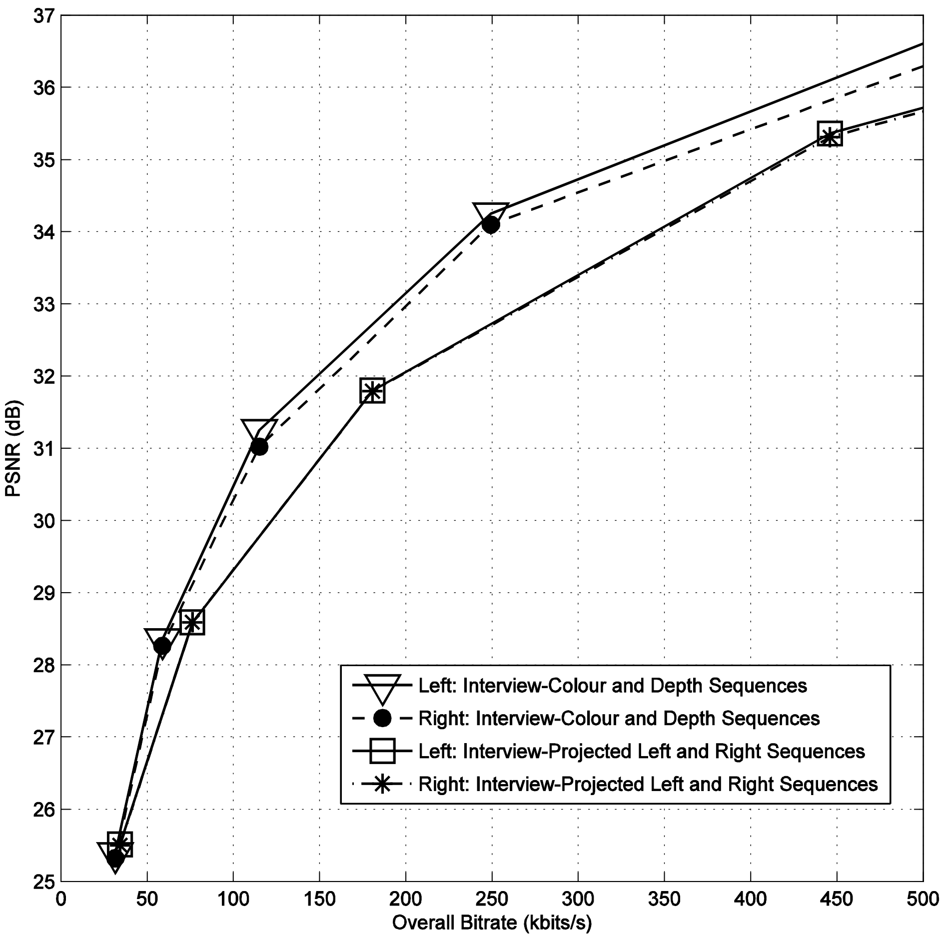


Figure 3.21: R-D curves for Interview (using colour and depth sequences and projected left and right sequences).

4 The transmission aspects of 3D video

Your goals for this "The transmission aspects of 3D video" chapter are to learn about:

- 3D video broadcasting.
- 3D video transmission over IP networks and 3D video streaming.
- 3D video error resilient and concealment techniques.
- Challenges for 3D video over unreliable networks.

With the availability of well defined scene representations, efficient coding approaches and affordable 3D displays, a need for efficient 3D video transmission technology has become acute. Unlike 3D movies, which can be traced back to 1903, 3D broadcast and streaming/communication applications require rate-adaptive coding approaches, transport streams for delivery and signalling (e.g. MPEG-2 Transport Stream (TS)), control protocols and error resilient tools to render good quality 3D video content to the end-users [60]. The practical demonstrations (e.g. 3D video streaming) and prototype 3D video transmission systems have being studied in the past [61-63]. However, the technology is not matured up to the level of 2D video transmission technologies as yet. Moreover, most of the 3D delivery techniques will be built on top of the existing infrastructure and tools for conventional video applications. This section presents the background related to potential 3D transmission technologies and supportive technologies (e.g. error recovery) necessary for efficient transmission of 3D video content over broadcast and communication links.

3D-TV broadcasting has been an interesting application scenario from the early ages of analogue TV [64] [65]. However, these services were not continued for a long time due to the low quality pictures visible with different viewing aids (e.g. polarized glasses) and requirements for specific display systems [65]. With the introduction of digital transmission technologies, the possibilities of broadcasting 3D-TV pictures are further studied and necessary technologies (e.g. MPEG-2 Multi View Profile) are developed. For example, the integration of stereoscopic TV with HDTV is studied in [66] and [67] using separate transmission of left and right views and side-by-side arrangement (frame packing format) of left and right image sequences respectively. These approaches utilize existing infrastructures and technologies for HDTV transmission system and thus not so efficient due to the limitations (e.g. resolution, bandwidth) imposed for 3D video. For instance, frame compatible formats yield low resolution for left and right images (both left and right images are horizontally down-sampled). Due to the flexibility of rendering interactive stereoscopic video, colour plus depth representation has been tested over broadcasting channels [68] [24]. The approach in [68] utilizes the "private" or "user data" of the MPEG-2 Transport Stream to send the depth information [69], whereas the multiplexing of colour plus depth within the MPEG-2 Transport Stream is studied in [24]. Moreover, the latter approach is now standardised as an amendment ISO/IEC 13818-1 (MPEG-2 Transport Stream) [70].

The availability of voice over IP (e.g. VoIP), TV pictures over IP (e.g. IPTV) has influenced 3D video services to select the IP packet network as one of the main transport mediums. The wide spread usage of IP to deliver video services over wired/wireless channels will enable 3D video over large application space. Moreover, the flexible integration of IP transmission aspects with the other layers of the protocol stack allows more space for adapting different 3D scene representations and optimum coding methodologies. The 3D streaming services are classified into four main categories in [71] namely;

- Server unicasting to a single client
- Server multicasting to several clients
- Peer-to-peer (P2P) unicasting
- P2P multicasting.

The emerging 3D video streaming applications would prefer RTP/DCCP/IP protocol [72] over conventional RTP/UDP/IP protocol due to effective video rate adaptation for streaming applications by DCCP (Datagram Congestion Control Protocol) to match the TCP-Friendly Rate Control (TFRC) [73]. The rate adaptation strategies for stereo and multi-view video offer more flexibility than the conventional video due to the availability of more than one view during transmission. For example, the rate for the stereoscopic video application can be adapted by choosing different resolutions for the right image sequence together with full-resolution left image sequence according to the binocular suppression theorem[52]. Several studies can be found for open loop and closed loop rate adaptation schemes for 3D video over UDP and DCCP protocols [74-76]. For example, the client-driven multi-view video streaming described in [75] reduces the bandwidth need through selecting only a small number of views for transmission based on the user's head position.

Interactive 3D video streaming will enable seamless, more involving and adaptable delivery of 3D content to end users. However, 3D video streaming over band-limited and unreliable communication channels can introduce artifacts on the transmitted 3D content. The effect could be much more significant compared to conventional 2D video streaming. For instance, the nature of 3D video source format (e.g. colour plus depth images vs. left and right views) and the way our Human Visual System (HVS) perceives channel introduced artifacts in 3D video is different from 2D video; as an example, colour plus depth map 3D video presentation may have to utilize impaired depth map information at the receiver-side to render novel views.

Video streaming over the Internet has become one of the most popular applications and Internet 3D video streaming is expected to become more popular in the future, also thanks to the recently standardized wireless systems, including WIMAX, 3GPP LTE / LTE advanced, the latest 802.11 standards, and advanced short range wireless communication systems, enabling the transmission of high bandwidth multimedia data. For such applications the target of the system design should be the maximization of the final quality perceived by the user, or Quality of Experience (QoE), rather than only of the performance of the network in terms of "classical" quality of service (QoS) parameters such as throughput and delay. 3D video services, and in particular those delivered through wireless and mobile channels, face a number of challenges due to the need to handle a large amount of data and to the possible limitations due to the characteristics of the transmission channel and of the device. This can result in perceivable impairments originated in the different steps of the communication system, from content production to display techniques.

The overall enjoyment or annoyance of 3D video streaming applications or services is influenced by several factors such as human factors (e.g., demographic and socio economic background), system factors (e.g., content and network related influences) and contextual factors (e.g., duration, time of the day and frequency of use). The overall experience can be analysed and measured by QoE related parameters which quantify the user's overall satisfaction about a service [110][111]. Quality of Service (QoS) related measurements only measure performance aspects of a physical system, with main focus on telecommunications services. Measuring QoS parameters is straightforward since objective, explicit technological methods can be used, whereas measuring and understanding QoE requires a multi-disciplinary and multi-technological approach. The added dimension of depth in 3D viewing influences several perceptual attributes such as overall image quality, depth perception, naturalness, presence, visual comfort, etc. For instance, an increased binocular disparity enhances the depth perception of viewers, although in extreme cases this can lead to eye fatigue as well. Therefore, the overall enjoyment of the 3D application could be hindered by the eye strain experienced by the end user. The influence of these attributes on the overall experience of 3D video streaming users is yet to be investigated.

The effect of transmission over band-limited and unreliable communication channels (such as wireless channels) can be much worse for 3D video than for 2D video, due to the presence in the first case of two channels (i.e., stereoscopic 3D video) that can be impaired in a different way; as a consequence the 3D reconstruction in the human visual system may be affected. Some networks introduce factors directly related to temporal domain de-synchronization issues. For instance delay in one view could lead to temporal de-synchronization and this can lead to reduced comfort in 3D viewing. The methods employed to mitigate these artifacts (e.g., error concealment) need to be carefully designed to suit 3D video applications. The simple application of 2D image/video methods would not work effectively in this case, as discussed in [112] for different error concealment algorithms for 3D video transmission errors. In [112] it is observed that is some cases switching back to the 2D video mode is preferred to applying 2D error concealment methods separately for left and right views to recover missing image information during transmission. There could be added implications introduced by these artifacts into our HVS. Therefore artifacts caused as a result of 3D video streaming can be clearly appreciated only by understanding how our HVS perceives different 3D video artifacts. Frame freezing mechanisms employed to tackle missing frames caused by transmission errors or delay could lead to temporal de-synchronization where one eye sees delayed content compared to the other eye. There are two implications associated to the case where one view is affected by transmission impairments:

- Binocular suppression
- Binocular rivalry

Our HVS is still capable to align and fuse stereoscopic content if one view is affected by artifacts due to compression, transmission, and rendering. Binocular suppression theory suggests that in these situations the overall perception is usually driven by the quality of the best view (i.e., left or right view), at least if the quality of the worst view is above a threshold value. However this capability is limited and studies show that additional cognitive load is necessary to fuse these views [113]. Increased cognitive load leads to visual fatigue and eye strain and prevents users from watching 3D content for a long time. This directly affects user perception and QoE. If one of the views is extremely altered by the transmission system, the HVS will not be able to fuse the affected views, and this causes binocular rivalry. This has detrimental effects on the final QoE perceived by the end user. Recent studies on 3D video transmission [112] have found that binocular rivalry is causing the overall perception to be affected and this effect prevails over the effect of binocular suppression. To avoid the detrimental effect of binocular rivalry, the transmission system could be designed appropriately taking this issue into account. For instance, the transmission system parameters can be updated "on the fly" to obtain 3D views with minimum distortions, according to the feedback on the measure of 3D video quality at the receiver-side. In case of low quality due to different errors in the two views, if the received quality of one of the views is significantly low, the transmission system could be informed to allocate more resources to the worse view or to increase the error protection level for that 3D video channel to mitigate the quality loss in the future. This increases the opportunity to fuse the 3D video content more effectively and improve the final QoE of users.

Video over IP often suffers from packet losses. The congestion and noise/interference/fading are the main causes for packet losses in wired and wireless links respectively. Therefore, effective error correction approaches (e.g. Automatic Repeat-reQuest (ARQ), Forward Error Correction (FEC)), joint source channel coding techniques and error concealment techniques are necessary to send 3D services over IP networks more effectively. The performance of 3D video transmission over wireless channels which is considered as bandwidth limited and error prone, can be further improved with Joint Source Channel Coding (JSCC) approach which is an effective method to overcome such challenges [77]. The study carried out in [78] proposes a JSCC scheme for colour plus depth stereoscopic video. The advanced channel coding approaches for 3D video are addressed in several studies [79]. For example, in [80], stereoscopic video streaming using FEC techniques are investigated. The left and right image frames are classified into layers based on their contribution towards final reconstructed quality. Then the layers are transmitted based on the principal of Unequal Error Protection (UEP) using different error correction codes. Furthermore, Multiple Description Coding (MDC) approaches are proposed for stereoscopic video which allows sending video with acceptable quality bounds at bad channel conditions [81]. However, the use of error correction codes with 3D video is somewhat unjustifiable due to the high demand for bandwidth by 3D video content itself. Therefore, this research explores different perspectives to protect 3D video over networks without sending additional data. For example, transmission power can be allocated unequally for 3D video components depending on their contribution towards perceptual quality. In a way the adapted approaches exploit cross-layer design using the perceptual aspects of 3D video.

Error concealment is necessary to perform at the decoder in order to reduce the temporal error propagation caused by unpreventable packet losses. The conventional error concealment tools for 2D video can be adapted in recovering the errors of corrupted 3D video [82][83]. Error concealment algorithms have been proposed for left and right view based stereoscopic video utilizing the additional data from the corresponding image sequence [84] [85]. Bilen et al propose two methods for full frame loss concealment in left and right based stereoscopic video using overlapped block motion and disparity compensation [86]. Moreover, in [87] an approach to recover the entire right frame in stereoscopic video transmission is described based on the relativity of prediction modes for right frames. However, the above mentioned studies are aimed at providing error resilience for left and right view based stereoscopic video. Therefore, error concealment of colour plus depth stereoscopic video is considered in this book. Three error concealment methods are proposed based on the correlation of scene characteristics (e.g. motion correlation) and shape concealment approaches.

5 3D video display technologies

Your goals for this "3D video display technologies" chapter are to learn about:

- 3D display categories.
- 3D viewing aids.
- Auto-stereoscopic displays.

3D displays are the last node in the 3D video chain and expect to produce pleasing pictures to end-users. The 3D displays generate a number of physiological cues such as binocular disparity, motion parallax, ocular convergence and accommodation in order to produce depth sensations in the Human Visual System (HVS). However, conflicting clues may cause discomfort and fatigue. Unlike standard video displays (e.g. PC monitor), 3D video requires special display techniques or additional viewing aids such as polarization glasses and shutter glasses. Furthermore, only the cost effective and adaptive displays will enable 3D video in every household. For, example rather than having a specific 3D display at home, a single display that can work as 2D or 3D display get more attention from the users. A survey of 3D display techniques is presented in [88]. The major types of 3D video display techniques and the display type used in this research are discussed in the following.

The 3D display techniques can be categorized into three main types namely;

- Volumetric displays

Figure 5.1: Volumetric display

- Holographic displays

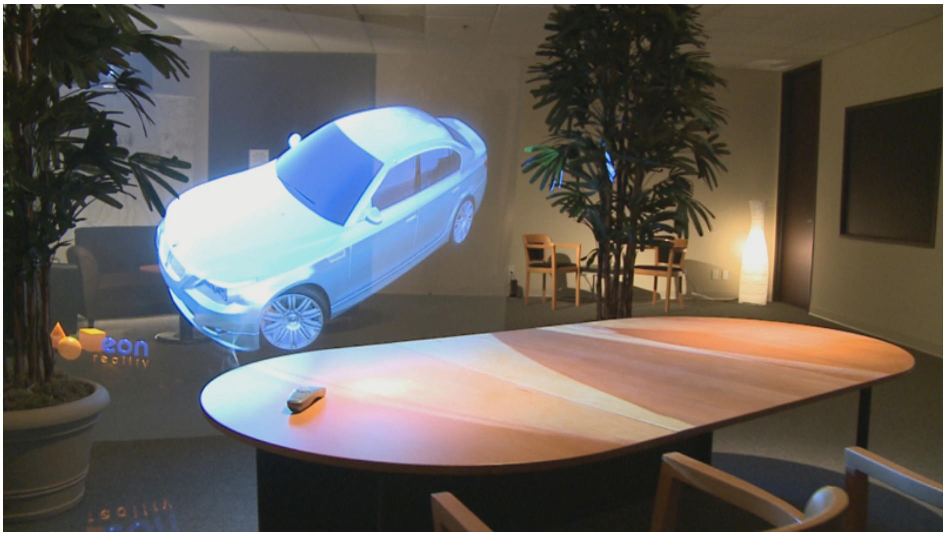


Figure 5.2: Holographic display

- Autostereoscopic displays

Figure 5.3: Auto-stereoscopic display

Volumetric displays form the image by projection onto a volume space. This allows the users to look at the scene from wide range of angles and viewpoints. However, these displays tend to have limited resolution. The descriptions of commercially available and prototype volumetric displays can be found in [89–91]. Due to the unavailability of natural scene capture for these displays, volumetric displays are more suitable for computer graphics than 3D video applications. In holographic display techniques, the image is formed by wave-front reconstruction, and includes both real and virtual image reconstruction. The quality of 3D video produced by colour holography is more rich in depth cues than the 3D video reproduced by other techniques [88] [92]. Some of the proposed display techniques using holography are detailed in [93-95]. The advantage of this technique is that it captures the true 3D wavefront of a scene and the retention of motion parallax. However, due to the demand for high resolution recording medium and replay medium, this display technique is difficult to implement for natural scenes with acceptable image resolutions with the existing technologies. For instance, a large display of say 100 mm diagonal will need dramatic improvements in Very Large Scale Integration (VLSI) techniques to enable a Spatial Light Modulator (SLM) to be manufactured with sufficient pixel resolution [88]. Nonetheless, holographic display technique can be deployed in reduced parallax systems (e.g. stereo-holography or lenticular), which reduce the demand for high-end technologies.

Autostereoscopic displays provide depth sensation to the viewers with no requirement for special glasses or user-mounted devices. Prior to the introduction of these displays viewers have to wear special viewing aids to facilitate stereoscopic viewing. The viewing methods can be classified into several categories depending on how they are going to render left and right image sequences to the user. According to [96], some of the widely used stereoscopic viewing methods are

- Optical (e.g. the stereoscope),

Figure 5.4: Stereoscope

- Colour based (e.g. anaglyphs),

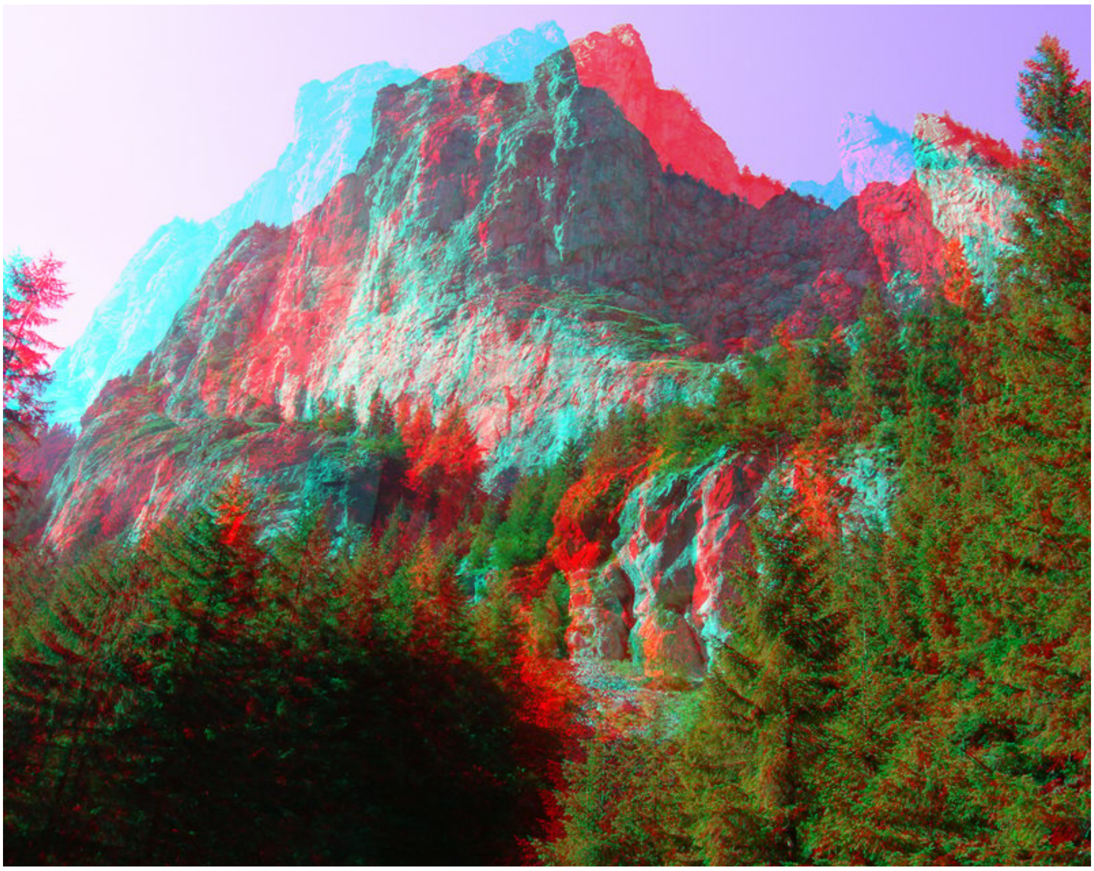


Figure 5.5: anaglyphs image (colour based)

Figure 5.6: anaglyphs glasses (Red-blue glasses)

- Polarization based (e.g. polarization glasses),

Figure 5.7: Polarized glasses (Red-blue glasses)

- Temporal separation based (e.g. liquid crystal shutter glasses), and

Figure 5.8: Shutter glasses

- Head Mounted Displays (HMD)

Figure 5.9: Head Mounted Display

HMD in particular renders the left and right views to the user based on their head position. These are mostly used in interactive 3D video applications. Even though the viewing 3D with eye-wears is cost effective solution, the users at home may not like to wear them due to the general discomfort. Moreover, these eye-wears introduce specific artefacts to the stereoscopic image sequences. For example, with polarization glasses each eye sees a faint version of the other eye's image [96]. Therefore, autostereoscopic displays provide convenient solution for 3D viewers and due to the wide popularity of this technique the cost for a display unit going to be lowered in the future.

These displays can be employed to produce binocular (single user), multi-view (multiple discrete stereoscopic views) and holoform views. Unlike other 3D displays (e.g. volumetric display) this display supports limited number of viewers. Two design approaches for autostereoscopic displays are shown in Figure 5.10. Figure 5.10 (a) shows a lenticular array of cylindrical lens-lets placed in front of the pixel raster, directing the light from adjacent pixel columns to different viewing slots at the ideal viewing distance. Therefore, with this approach, each eye of the viewer sees light from only every second pixel column and thus provides stereoscopy. Figure 5.10 (b) shows a Parallax barrier in which a mask is placed in front of the pixel raster so that each eye sees light from only every second pixel column. These autostereoscopic displays divide the horizontal resolution of the underlying, typically liquid-crystal display device into two sets. One of the two visible images (e.g. left view) consists of every second column of pixels; the second image (e.g. right view) consists of the other columns. The two image sequences (i.e. stereoscopic video) are captured or generated with the DIBR method so that one is appropriate for each of the eye pair [97]. As this technology is still at the early age of development, there are certain issues to be addressed such as flipping between views and vertical binding on the image [88].

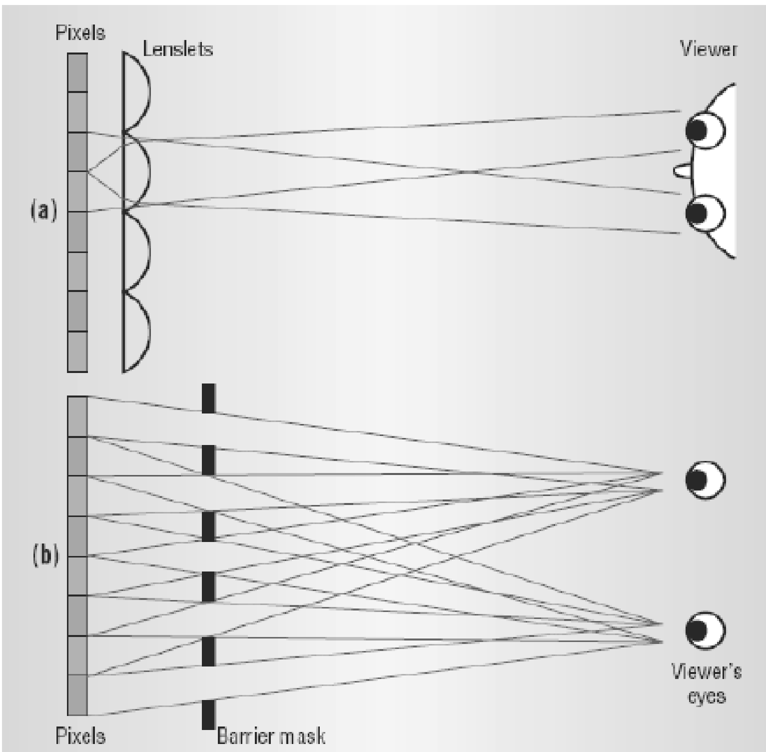


Figure 5.10: Auto-stereoscope arrangements

Autostereoscopic displays are commercially available as Mobile/tablets/PC/laptop monitors [98], and TV monitors [99]. Philips and Sharp Corporations have also developed Autostereoscopic displays which can work with both 2D and 3D content [98][99]. The 42" Philips multi-view Autostereoscopic display is used in the experiment to display the stereoscopic material. This exploits lenticular lens technology to separate left and right views [99]. The maximum resolution of the 3D display available is 1920×1080 pixels and the optics are optimized for a viewing distance of 3 meters. Nine users are allowed to view 3D video content at the same time. Moreover, this display supports screen parallax enabling users to look around the scene objects (see Figure 5.11). The input video format of this display is colour and depth map video sequences, which are arranged side-by-side. Therefore, the processed/original colour plus depth image sequences are combined to form side-by-side image sequences before fed them to the display. Before displaying 3D content on the display, the supplied colour and depth image sequences are converted into left and right image sequences using the DIBR technique.

Figure 5.11: Viewing of stereoscopic content using multi-view display

6 Quality evaluation of 3D video

Your goals for this "Quality evaluation of 3D video" chapter are to learn about:

- Perceptual quality evaluation of 3D video.
- Subjective 3D video quality evaluation methods.
- Objective 3D video quality evaluation methods.
- Real-time 3D video quality evaluation methods.

Even though the initial developments for 3D video services are in place, the acceptance of these services is dependent on the user satisfaction of the reconstructed 3D video quality. Therefore, extensive quality evaluation studies are necessary to study the effect of camera arrangement, data representation, coding, transmission and display techniques on the perceived quality of 3D video. Some of the stereoscopic image impairments introduced by the 3D video system are keystone distortion, depth-plane curvature, crosstalk, size distortions, cardboard effect, picket fence effect, image flipping and shear distortion. Moreover, depending on the coding approaches (e.g. DCT) being used, conventional coding impairments like blockiness, blur will be introduced to the reconstructed 3D video. These impairments in stereoscopic video will influence multi-dimensional perceptual attributes such as image quality, depth perception, presence, naturalness, etc. A detailed analysis is necessary to study how these 3D percepts influence the overall perceived quality in general. For instance, the study presented in [100] concludes that excessive disparities can cause eye strain and therefore degrade the perceived image quality. Mostly psychophysical experiments are conducted to measure and quantify 3D perceptual attributes. In addition to that, explorative studies can be utilized to get unbiased attitudes and views for emerging technologies like 3D video. For instance, focus groups can be formed to evaluate the impact of new stereoscopic image systems through group discussions [101]. This method also can be employed to evaluate the added value of depth. Moreover, explorative studies will help better understanding the attributes of a multi-dimensional construct like image quality, depth perception, viewing experience, etc.

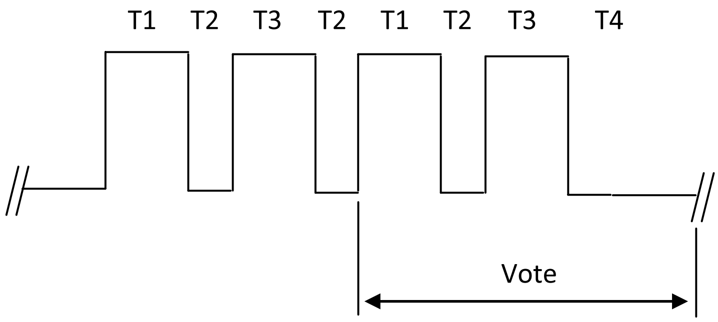
Psychophysical scaling paradigms can be classified into two main categories [102], namely;

- Performance-oriented methods
- Appreciation oriented methods.

The performance oriented assessment methods are utilized to measure the effectiveness of a specific task whereas appreciation oriented methods measure and quantify the perceptual attributes of new media types and decide whether the content is pleasing or not. Appreciation oriented quality evaluation methodology for stereoscopic TV pictures is described in ITU-R BT.1438 recommendation [103]. Most of the subjective evaluation procedures in this recommendation are based on the ITU quality evaluation recommendation for TV pictures (i.e. ITU-R BT.500.11) [104]. In addition to the measurement of image quality, other 3D perceptual attributes like presence, naturalness and eye strain can be measured using the same experimental paradigms. The main evaluation strategies mentioned in [103] are;

- Single-Stimulus-Continuous-Quality-Scale (SSCQS) method: The quality is assessed individually for each stereoscopic image sequence in the stimulus set.
- Stimulus comparison method: Series of stereoscopic image sequences are presented sequentially in time and observers are asked to assign a relation between two consecutive stereoscopic video sequences
- Double-Stimulus-Continuous-Quality-Scale (DSCQS) method: Alternately, an unimpaired stereo image sequence (reference) and an impaired stereo image sequence (test) are shown. The reference and test image sequences are presented twice. For both stereo image sequences (reference and test) observers assess the overall picture quality separately.

The presentation method of DSCQS is illustrated in Figure 6.1. The final quality rating (i.e. opinion score) of this method is the difference of individual scores for the reference and impaired image sequences. Subsequently the individual opinion scores are averaged across all the subjects in order to obtain the Mean Opinion Score (MOS). The confidence intervals can also be specified to indicate the individual differences.



Phases of presentation:

T1 = 10 s	Test sequence A
T2 = 3 s	Mid-gray
T3 = 10 s	Test sequence B
T4 = 5-11 s	Mid-gray

Figure 6.1: DSCQS presentation structure

The DSCQS and SSCQS methods are utilized in most the experiments described in this book as main subjective quality evaluation methods as these methods are recommended by standardization bodies for stereoscopic video quality measurements and are in wider usage in 3D video research [103], [105-107]. Furthermore, all subjects are screened for their visual acuity (using the Snellen chart), good stereo vision (using the TNO stereo test), and good colour vision (the Ishihara test). Moreover, the 3D displays will be calibrated using the GretagMacbeth Eye-One Display 2 calibration device and test environments (e.g. home viewing conditions) will be set according to the specifications set by ITU-R BT.500.11 recommendation.

The perceptual quality of asymmetrically coded colour and depth map sequences are measured and evaluated in [130]. Moreover, the effect of packet losses on the perceptual quality is also studied. The quality is measured across two perceptual attributes namely, image quality and depth perception. Furthermore, the relationships are derived among these measured perceptual attributes.

Subjective tests for each 3D video system parameter (e.g. camera angle, coding) change is not an efficient method to evaluate the quality due to several reasons. The most prominent reasons are the time consumption, enormous effort necessary, and the requirements for special test environments (e.g. standard test laboratories). Therefore, candidate objective quality measures of 3D video have become a compromise way of measuring the quality.

Therefore, candidate objective quality measures (i.e. PSNR) of colour image sequence and depth image sequence are utilized to represent the effectiveness of proposed algorithms in this book. PSNR is derived by setting the Mean Squared Error (MSE) in relation to the maximum possible value of the luminance (see Equations 6.1 and 6.2).

For n-bit value this is as follows,

$$MSE = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N}\left[g(i,j)-G(i,j)\right]^2}{M \cdot N}$$ Equation 6.1

$$PSNR = 20 \cdot \log_{10}\left(\frac{2^n - 1}{\sqrt{MSE}}\right)$$ Equation 6.2

Where g (i,j) is the original signal at pixel (i,j), G (i,j) is the processed signal and M × N is the picture size. The resultant is a single number in decibels (dB).

Even though PSNR scores of depth image are indicative, it may not represent the depth as perceived by the human observers. Therefore, the objective quality measures of rendered left and right views using the DIBR method are also used to quantify the depth perception. In order to obtain PSNR ratings the left and right video rendered using the impaired colour and depth image sequences are compared against the left and right video rendered using the original/reference colour and depth map sequences.

These objective measures may or may not strongly correlate with the quality attributes of 3D video as measured by subjective tests. Studies have found out that there is a high correlation between subjective ratings and individual objective quality ratings of 3D video components (e.g., average PSNR and SSIM of left and right video or colour and depth video) [123]. For instance, depth perception is highly correlated to the average PSNR of the rendered left and right image sequences [123]. This could be due to the loss of correspondence between left and right objects and reduction of monocular depth cues as a result of compression and transmission errors. This means that we could use individual objective quality measures of different 3D video components to predict the true user perception in place of subjective quality evaluation, through a suitable approximation derived based on correlation analysis. However, with some 3D source representations such as the colour and depth map 3D image format, it may be difficult to derive a direct relationship between objective measures and subjective quality ratings. For instance, the objective quality of the depth map may have a very weak correlation on its own with the overall subjective quality, because the depth map is used for projecting the corresponding colour image into 3D coordinates and it is not directly viewed by the end users. Individual quality ratings of left and right views may not always account for depth reproduction of the scene. Therefore, the next phase of 3D objective quality metrics includes a methodology to quantify the effect of binocular disparity of 3D scenes in addition to a conventional image/video quality assessment methodology. For instance in [122], in addition to image quality artifacts, disparity distortion measures were also incorporated to evaluate the overall 3D video quality. The article showed improved performance over the method which does not account for the correspondence information of stereoscopic views. The latest 3D image/video quality metrics evaluate depth reproduction in addition to usual image artifacts (such as blockiness) using specific image features (e.g., edge, disparity and structural information of stereoscopic images) which are important for the HVS in both 2D and 3D viewing. For instance the method proposed in [124] shows high correlation values with subjective quality results (Mean Opinion Score, MOS): the correlation coefficient with subjective quality ratings is as high as 0.95; this outperforms the method based on 2D image quality + disparity [122] and other conventional 2D quality metrics separately applied to left and right views (see Table 6.1). The reported performance figures in Table 6.1 are obtained using the same 3D dataset. These observations confirm that accurate 3D image quality metrics should be designed to also consider binocular disparity distortions. All the methods described above are Full-Reference (FR) methods and need the original 3D image sequence to measure the quality by comparison, hence they are not suitable for the evaluation of the quality "on the fly" in real-time transmission applications such as interactive 3D video streaming. In this case the solution is to use Reduced-Reference (RR) or No-Reference (NR) metrics which do not require the original image for quality assessment, but either no information (NR) or just some side-information about it (RR) requiring few bits for its transmission. Most of the NR metrics are designed specifically for a known set of artifacts (e.g., JPEG compression) and cannot be deployed in a more general scenario. In case of RR metrics, side-information is generated from features extracted from the original 3D image sequence and sent to the receiver-side to measure 3D video quality. Since the reference side-information has to be transmitted over the channel, either in-band or on a dedicated connection, the overhead should be kept at a minimum level. The next section describes how we could measure 3D video quality "on the fly" using RR and NR methods and provides a brief description of the existing methods.

Method	CC	SSE	RMSE
SSIM (Structural SIMilarity)	0.837	0.965	0.159
VQM (Video Quality Metric)	0.932	0.423	0.106
Proposed in [8]: 2D image quality + Disparity	0.901	0.608	0.126
Proposed in [10]	0.947	0.341	0.095

Table 6.1: Correlation between objective 3D image/video measures and subjective quality

6.1 Real-time 3D video quality evaluation strategies

The measured image quality at the receiver-side can be used as feedback information to update system parameters "on the fly" in a "QoE-aware" system design approach [117][125]. However, measuring 3D video quality in real time is a challenge mainly due to the complex nature of 3D video quality and also the fact that the amount of side-information to be sent to measure the quality with RR methods is larger compared to 2D image/video applications. The emerging RR and NR quality evaluation methods are based on image features associated to the characteristics of the HVS. Some of these features are related to image perception (e.g., luminance, contrast) and some are related to depth perception (e.g., disparity, structural correlations). An appropriate selection of these features is crucial to design an effective 3D image/video quality assessment method. The selected features should be able to quantify image and depth perception related artifacts with a minimum overhead. If the overhead is significant, the feasibility of deploying the designed RR method is reduced. Figure 6.2 shows how the extracted edge information is employed to measure 3D video quality in the RR method proposed in [126]. In this method, luminance and contrast details of the original and distorted images are utilized to count for conventional image artifacts, whereas edge information based structural correlation is employed to measure the structural/disparity degradation of the 3D scene, which is directly affecting rendering using colour plus depth map based 3D video. In order to reduce the overhead for side-information (i.e., extracted features of the reference image) lossless compression mechanisms can be deployed for its compression. An extra effort should be also made to send the side-information without corruption using a dedicated channel or highly protected forward channel. Visual attention models could also be utilized to find 3D image/video features which attract significant attention during 3D viewing. However, a direct relationship between visual attention and image perception for 3D images and video is yet to be found. NR methods are the most suitable for real-time 3D video applications since these do not consume any bandwidth for the transmission of side information. However, their performance and application domain is limited since they rely solely on the received 3D image/video sequence and other contextual information (e.g., Hybrid-NR methods: packet loss rate, bit-error rate). It may be impossible to count for all the artifacts imposed along the end to end 3D video chain without referring to the original image sequence. This is why most of the proposed NR metrics are limited to a specific set of artifacts [127].

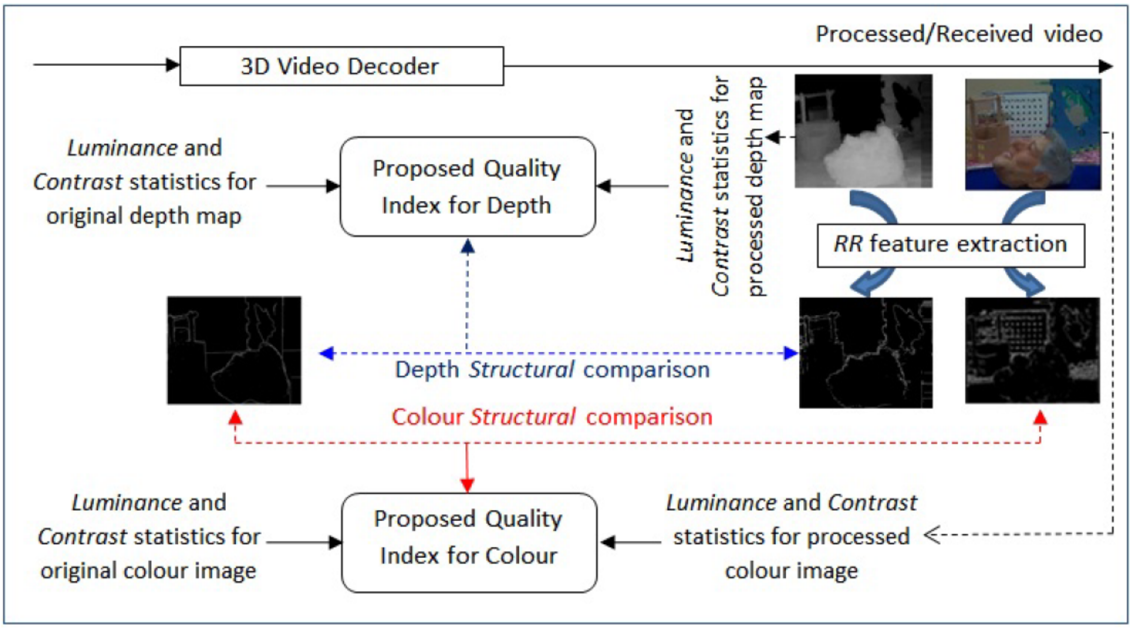


Figure 6.2: Reduced reference edge based 3D video quality metric [12].

Table 6.2 reports a few existing NR and RR quality metrics for 3D image/video. This table explains which image features are used to measure the overall perception and how much the different metrics are correlated with subjective quality scores (i.e., MOS) and with existing Full-Reference methods. It can be observed that most of these methods show a high degree of correlation with subjective MOS and Full-reference methods. However, these metrics are focused on one or two specific 3D perceptual attributes. The combined effect of these perceptual attributes which is directly related to user 3D QoE has not been addressed to date. The methods in [128] and [127] are evaluated using the same image database whereas others are evaluated using different data sets. Since some of these metrics, e.g., NR metrics ([127] and [129]) are designed for a particular types of image artifacts (e.g., JPEG compression), it is not always possible to compare the performance of a NR metric with another objective quality model in a common dataset. On the other hand, due to the overhead associated with RR metrics compared to zero overhead for NR metrics, the usage and advantages of these methods are significantly different. In addition, due to some practical reasons (intellectual property rights, different source 3D video formats, e.g., colour + depth vs. left and right images, unavailability of ground truth depth maps, etc.), it is not always feasible to compare the performance of two different 3D quality evaluation algorithms in a common dataset. The lack of reliable and comprehensive 3D image/video databases is another major challenge faced by researchers and developers, making difficult to effectively compare the performance of emerging objective and subjective quality evaluation methods with that of the existing methods.

Quality Metric	Method (NR or RR)	Artifacts	Features used to measure image artifacts (IA) and disparity (D)	CC	ROCC	OR	RMSE
Cyclop [14]	RR	JPEG symmetric and asymmetric coding artifacts	IA: Contrast sensitivity (spatial frequency and orientation); D: coherence of cyclopean images	0.981	0.950	0.050	-
Sazzad et al. [13]	NR	JPEG symmetric and asymmetric coding artifacts	IA: Blockiness and zero crossing of edge, flat and texture areas; D : average zero crossing of plane and non-plane areas	0.960	0.920	0.069	-
Solh et al. [15]	NR	Depth map and colored video compression, depth estimation (stereo matching), and depth from 2D to 3D conversion	IA D: Temporal outliers (TO), temporal inconsistencies (TI), and spatial outliers (SO) using ideal depth estimate for each pixel	0.916	0.1003	0.8	1.686
Hewage & Martini [12]	RR	H264 compression and random packet losses	IA: Luminance, structure and contrast D : edge based structural correlation	Colour: 0.9273 (vs. FR); Depth: 0.9795 (vs. FR)			Colour: 0.0110 (vs. FR); Depth: 0.0064 (vs. FR)

Table 6.2: No-Reference (NR) and Reduced-Reference (RR) methods for 3D image/video

6.2 Challenges for real-time 3D video quality evaluation

The possibility to measure 3D image/video quality in real time, as requested by 3D video applications, is hindered by several issues. The major challenge is how we could measure the effect of all perceptual attributes (e.g., depth, presence, naturalness, etc.) associated with 3D viewing. The lack of availability of common 3D image/video databases is also detrimental for the advance in this discipline. The following paragraphs briefly discuss these challenges and possible solutions foreseen.

6.2.1 Measurement of different 3D perceptual attributes

Even though emerging 3D quality evaluation methods accurately predict a given quality attribute, the relationship among these perception attributes has not be thoroughly studied. The combined effect directly affects user experience and can be measured using emerging QoE indices. Therefore the current need is to understand how 3D audio/image processing and transmission artifacts affect the overall experience of the user, then identify audio, image and contextual features which can be used to quantify the overall effect on user experience. On the other hand, it is necessary to understand how the HVS perceives these 3D artifacts. For instance, there could be conflicts based on whether binocular suppression or binocular rivalry is taking place based on the artifacts in question. These aspects need extended attention in order to measure the overall experience of 3D viewing. In order to enable a unified approach to 3D objective quality subjective quality evaluation studies, standardization of these procedures are necessary. Several standardization activities are being carried out by VQEG, ITU (Recommendations: ITU-T P- and J-series), European Broadcasting Union EBU (3D-TV Group) and other Standards Developing Organizations (SDOs) in relation to 3D video subjective and objective quality evaluations. Currently, the Video Quality Expert Group (VQEG) is working (3DTV project) on creating a ground truth 3D video dataset (GroTruQoE dataset) using the pair-comparison method. This ground truth database will then be used to evaluate other time-efficient 3D subjective quality evaluation methodologies and objective quality models. In addition, the project also addresses the objective quality assessment of 3D video, with the plan to evaluate 3D quality of experience in relation to the visual quality, depth quality and visual comfort dimensions. Most of these findings are reported to objective and subjective 3D video quality studies in ITU-T Study Groups (SG) 9 and 12. EBU is also working on 3D video production, formats and sequence properties for 3DTV Broadcasting applications (e.g., EBU Recommendation R 135).

B. Lack of 3D image/video databases

There are several image/video quality databases for conventional 2D image/video artifacts, although only a few have been reported for 3D image/video artifacts. This prevents developers from using a common dataset to evaluate the performance of their metrics. Table 6.3 shows some of the reported 3D image/video databases in the literature. The amount of artifacts considered in these databases is limited. Most of them do not consider artifacts which could be introduced during transmission. Therefore it is a responsibility of the research community to produce comprehensive 3D video datasets covering a range of image and transmission artifacts and make available the developed 3D image/video dataset publicly.

3D image/video database	Creator	Artifacts
Mobile 3D video database	University of Tampere and Nokia	Crosstalk, blocking, colour mismatch and bleeding, packet losses for low-resolution video (only impaired sequences, no MOS values provided).
IRCCyN 3D image database	University of Nantes	JPEG, J2K, upsample/downsample, etc.
EPFL databases for images/videos	EPFL	Different camera distances
Kingston University video database	Kingston University-London	Packet losses
NAMA3DS1-COSPAD1	University of Nantes	H.264 and JPEG2000 compression artifacts
RMIT3DV	RMIT University	Uncompressed HD 3D video

Table 6.3: Available 3D image/video database

C. Visual attention models to develop RR and NR quality metrics

The attention of users during 3D viewing can be influenced by several factors including spatial/temporal frequencies, depth cues, conflicting depth cues, etc. The studies on visual attention in 2D/3D images found out that the behaviour of viewers during 2D viewing and 3D viewing is not always identical (e.g., centre bias vs. depth bias). These observations are tightly linked with the way we perceive 3D video. Therefore, effective 3D video quality evaluation and 3D QoE enhancement schemes could be designed based on these observations. There are still unanswered questions such as whether quality assessment is analogous to attentional quality assessment and also how attention mechanisms could be properly integrated into design of QoE assessment methodologies. A thorough study has not been conducted to date in order to identify the relationship between 3D image/video attention models and 3D image/video quality evaluation. Similar to the integrated model described above, attentive areas identified by visual attention studies can be utilized to extract image features which can be used to design No-Reference (NR) and Reduced-Reference (RR) quality metrics for real-time 3D video application. Furthermore, since visual attention models can predict the highly attentive areas of an image or video, these can be integrated into source and channel coding at the sender side. Emerging 3D saliency models incorporate 2D image, depth and motion information which can be applied to 3D video sequences. Most of the reported 3D saliency models are extensions of 2D visual saliency models by incorporating depth information. Table IV summarises a few 3D saliency models reported in the literature. There are two main types of depth integrated saliency models, namely: Depth weighted 3D saliency model and Depth saliency model based methods. The depth weighted saliency models weight the 2D saliency map based on depth information. In depth saliency models, the predicted 3D saliency map is derived based on the chosen weights for 2D and depth saliency maps.

7 Conclusion

3-D video capture and display technologies have experienced rapid growth and commercial success during the last decade. Driven by the powerful vision of being able to communicate in 3-D, the integration of 3-D video and communication technologies are underway. Immersive communication applications provide more natural conditions for effective human interaction. 3-D video communication is demanding due to the stringent requirements on quality and the enormous amounts of data involved. 3D video has become ever closer to the users due to the technological advancement of related technologies. This book has viewed the state of the art of 3D video technologies, ranging from 3D video capture to display. Different 3D video scene representations and capturing techniques are presented and the advantages and disadvantages of the colour plus depth map representation, which is the selected representation for the experiments are discussed. Next, the compression techniques for stereoscopic video are described. This also elaborates on principles of classical 2D video coding approaches and scalable video coding. The transmission technologies are presented with more focus on 3D video over IP networks. Furthermore, this elaborates on error resilient and error concealments methods applicable for stereoscopic video. The viewing methods for 3D video are getting popular and becoming more affordable. The state of the art 3D video display techniques are discussed including latest Autostereoscopic display techniques. Finally, the quality issues of 3D video are explored and formal methodologies of evaluating emerging 3D video application scenarios are presented.

7.1 Areas for future research

This section describes some of the issues, which remain to be tackled in the provision of 3-D communication services and outlines the author's view of the future of 3-D multimedia communications.

Efficient encoding approaches are proposed for colour plus depth stereoscopic video applications in Chapter 3. The conventional block-based coding algorithms are utilized to encode both colour and depth image sequences in the proposed methods. Motion parameters for colour and depth map video are estimated using 2D motion search algorithms. However, 2D motion estimations will not account for all the motion in the depth images. For example, the motion in the direction of z-axis will not be tackled by the 2D motion estimation process. Therefore, dedicated 3-D motion estimation algorithms can be designed to get better compression efficiency than the use of 2D motion search algorithms.

Scalable video encoding approaches for 3D video as described in Chapter 3 can be effectively used in reducing the storage and bandwidth required for 3-D video applications without any degradation to perceived quality. This can be further extended to support bandwidth variations in the network adaptively. Closed-loop rate adaptation algorithms can be designed to optimize the 3-D video quality under changing network conditions. Furthermore, the asymmetric coding methods can be extended to design Joint Source Channel Coding (JSCC) methods. For instance, when channel conditions are bad, the depth image can be encoded with reduced resolution to allow for more channel protection to be applied to the 3-D stream. The full resolution depth images can be sent with reduced protection at good channel conditions. In case of a backward compatible 3-D video service, the transmitter can only send the colour video stream with full protection at bad channel conditions by neglecting the associated depth stream.

Even though coding of colour plus depth based stereoscopic video is considered in this book, the proposed techniques can also be applied to emerging multi-view plus depth map representation. Unlike stereoscopic video, the demand for system resources (e.g. storage and bandwidth) is enormous for multi-view video applications. Therefore, the proposed encoding approaches can be extended to design bandwidth friendly multi-view video applications.

Unequal error protection (UEP) method for backward compatible 3-D video transmission over WiMAX is proposed by the author in [133]. The allocation of transmission power is varied to support unequal error protection for colour plus depth video. Even though the quality is optimized in this approach the transmission power can also be optimized with the proposed method. Furthermore, this method can also be extended to multi-view video applications. For example, selected number of views can be highly protected based on the user's head position. In addition, the use of transmission power to provide unequal protection would be more suitable for multi-view applications, because adding redundancies for protection may not be the best way to provide protection due to the high bandwidth demand of these applications.

3-D video distribution in a home environment is also addressed by the authors in [132]. A prioritization scheme is proposed for 3-D video distribution over wireless network based on their importance towards better perceptual quality [132]. This work can be further extended to optimize 3-D video quality adaptively based on the type of network traffic in the premises. For instance, when there is not much video traffic, both colour and depth streams can be mapped to a high priority traffic class whereas depth stream is mapped to a lower priority traffic class compared to that of the colour stream when the network is flooded with video traffic. The multi-view plus depth video distribution can also be prioritized based on the proposed prioritization scheme. For example, the important views which are necessary to be able to render other views can be allocated a higher priority traffic class to optimize the rendered 3-D video quality.

Depth error concealment algorithm is proposed for colour plus depth stereoscopic video by the author in [131]. This method exploits the existing correlation between colour and depth map sequences. The motion correlation between colour and depth map frames may not be high at all time instances. Therefore, the proposed method can be joined with conventional frame concealment algorithms (e.g. frame copy) to achieve better results. Accurate depth maps are required to render novel views in multi-view plus depth 3-D video. Hence, the proposed method can be utilized to render good quality novel views for multi-view applications.

References

1. A. Smolic, and H. Kimata, Report on Status of 3DAV Exploration, *ISO/IEC JTC1/SC29/WG11*, Doc. no. N5558, Pattaya, Thailand, March 2003.
2. A. Smolic and H. Kimata, Applications and Requirements for 3DAV, *ISO/IEC JTC1/SC29/WG11 W5877*, Doc. no.W5877, Norway, July 2003.
3. Immersive media, http://www.immersivemedia.com/, 2014.
4. T.E. Boult, R.J. Micheals, M. Eckmann, X. Gao, C. Power, and S. Sablak, "Omnidirectional video applications", *Proceedings of the 8th International Symposium on Intelligent Robotic Systems*, 2000.
5. A. Smolic, and H. Kimata, Report on 3DAV Exploration, *ISO/IEC JTC1/SC29/WG11 W5878*, Doc. no. W5878, Norway, July 2003.
6. T. Fujii, and M. Tanimoto, "Multi-View Video Acquisition System for FTV Experiment", *ISO/IEC JTC1/SC29/WG11 N9472*, Doc. no. N9472, Pattaya, Thailand, March 2003.
7. A. Puri, R.V. Kollarits, and B.G. Haskell, "Basics of stereoscopic video, new compression results with MPEG-2 and a proposal for MPEG-4", *Signal Processing: Image Communication,* vol. 10, pp. 201–234, 1997.
8. The Optometrists Network, http://www.vision3d.com/stereo.html, 2014.
9. 3-D Media Cluster, Three-Dimensional Community, http://www.3diim-cluster.eu/, May 2014.
10. L.M.J. Meesters, W.A. IJsselsteijn, and P.J.H. Seuntins, "Survey of perceptual quality issues in three-dimensional television systems", *Proceedings of the SPIE* Stereoscopic Displays and Virtual Reality Systems X, vol. 5006, no., pp. 313–326, 2003.
11. A. Woods, T. Docherty, and R. Koch, "Image Distortions in Stereoscopic Video Systems", *Proceedings of the SPIE: Stereoscopic Displays and Applications IV*, vol. 1915, pp. 36–48, San Jose, CA, Feb. 1993.
12. Stereo3D.com, http://www.stereo3d.com/vidrec.htm, 2014.
13. M. Pollefeys, R. Koch. M. Vergauwen, and G.L. Van, "Hand-Held Acquisition of 3D Models with a Video Camera", *Proceedings of Second International Conference on 3-D Digital Imaging and Modelling*, pp. 14–23, October 1999.
14. O. Wilinski, and K. Van Overveld, "Depth From Motion Using Confidence Based Block Matching", *Proceedings of Image and Multidimensional Signal Processing Workshop*, pp. 159–192, July 1998.
15. P. Harman, J. Flack, S. Fox, and M. Dowley, "Rapid 2D to 3D conversion", *Proceedings of the SPIE: Stereoscopic Displays and Virtual Reality Systems IX*, vol. 4660, pp. 78–86, 2002.
16. Blue Box, http://www.philips.com/global/en/gmm/images/3d/3dcontentceationproducts/downloads/BlueBox_white_paper.pdf, 2014.
17. M. Kawakita, T. Kurita, H. Kikuchi, Y. Yamanouchi, S. Inoue, and K. Iizuka, "High-definition three-dimension camera: HDTV version of an axi-vision camera", *NHK Laboratories Note*, vol. 479, 2002.

18. G. Iddan, and G. Yahav, "3D imaging in the studio (and elsewhere…)", *Proceedings of the SPIE 4298*, vol. 4298, pp. 48–55, 2001.
19. M. Kawakita, K. Lizuka, T. Aida, H. Kikuchi, H. Fujikake, J. Yonai, and K. Takizawa, "Axi-vision camera (real-time distance-mapping camera)", *Applied Optics*, vol. 39, no. 22, pp. 3931–3939, 2000.
20. L. McMillan, "An Image-Based Approach to Three-Dimensional Computer Graphics", *PhD book*, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA, 1997.
21. W.R. Mark, "Post-Rendering 3D Image Warping: Visibility, Reconstruction, and Performance for Depth-Image Warping", *PhD book*, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA, Apr.1999.
22. R.I. Hartley, and A. Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, Cambridge, UK, 2000.
23. G. Xu, and Z. Zhang, *Epipolar Geometry in Stereo, Motion and Object Recognition*, Kluver Academic Publishers, Dordrecht, The Netherlands, 1996.
24. C. Fehn, "Depth-image-based rendering (DIBR), compression and transmission for a new approach on 3D-TV", *in Proc. SPIE Conf. Stereoscopic Displays and Virtual Reality Systems XI*, vol. 5291, pp. 93–104, CA, U.S.A., Jan. 2004.
25. P. Milgram, and M. KrAuger, "Adaptation Effects in Stereo Due to On-line Changes in Camera Configuration", *In Proc. of SPIE Stereoscopic Displays and Applications '92*, pp. 122–134, San Jose, CA, USA, Feb. 1992.

26. J. Shade, S. Gortler, L.W. He, and R. Szeliski, "Layered Depth Images," *in Proc. of ACM SIGGRAPH'98*, pp. 231–242, Orlando, FL, USA, July 1998.
27. C. Fehn, R. De La Barre, and S. Pastoor, "Interactive 3-D TV: Concepts and Key Technologies", *Proceedings of the IEEE*, vol. 94, pp. 524–538, 2006.
28. Text of ISO/IEC FDIS 23002-3 Representation of Auxiliary Video and Supplemental Information, *ISO/IEC JTC1/SC29/WG11 N8768*, Doc. N8768, Marrakesh, Morocco, Jan. 2007.
29. P. Merkle, A. Smolic, K. Muller, and T. Wiegand, "Multi-View Video Plus Depth Representation and Coding, *IEEE International Conference on Image Processing (ICIP'07)*, pp. 201–204, San Antonio, TX, Oct 2007.
30. C. Fehn, K. Schuur, I. Feldmann, P. Kauff, and A. Smolic, "Distribution of ATTEST test sequences for EE4 in MPEG 3DAV", *ISO/IEC JTC1/SC29/WG11 MPEG02/M9219*, Doc. no. M9219, Awaji Island, Dec. 2002.
31. C.L. Zitnick, S.B. Kang, M. Uyttendaele, S. Winder, and R. Szeliski, "High-quality video view interpolation using a layered representation", *ACM SIGGRAPH and ACM Trans. on Graphics*, pp. 600–608, Los Angeles, CA, Aug. 2004.
32. Information technology – Coding of audio-visual objects – Part 10: Advanced Video Coding, *ISO/IEC JTC 1/SC 29 14496-10:2008*, 2008.
33. T. Wiegand, and G.J. Sullivan, "Overview of the H.264/AVC Video Coding Standard", *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 560–576, July 2003.
34. J. Ostermann, J. Bormans, P. List, D. Marpe, M. Narroschke, F. Pereira, T. Stockhammer, and T. Wedi, "Video coding with H.264/AVC: Tools, Performance, and Complexity", *IEEE Circuits and Systems Magazine*, vol. 4, no. 1, pp. 7–28, First Quarter 2004.
35. T. Stockhammer, M. Miska, M. Hannuksela, and T. Wiegand, "H.264/AVC in Wireless Environments", *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 657–673, July 2003.
36. S. Wenger, "H.264/AVC over IP", *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 645–656, July 2003.
37. C.T.E.R. Hewage, H. Kodikara Arachchi, T. Masterton, A.C. Yu, H. Uzuner, S. Dogan, and A.M. Kondoz, "Content Adaptation for Virtual Office Environment Using Scalable Video Coding", *Proceedings of the 16th IST Mobile and Wireless Communications Summit (IST Summit'07)*, pp. 1–5, Budapest, Hungary, July 2007.
38. B. Pesquet-Popescu, and V. Bottreau, "Three-dimensional lifting schemes for motion-compensated video compression," *Proceedings of ICASSP*, vol. 3, pp. 1793–1796, May 2001.
39. N. Mehrseresht, and D. Taubman, "An Efficient Content-Adaptive MC 3D-DWT with Enhanced Spatial and Temporal Scalability", *IEEE International Conference on Image Processing (ICIP'04)*, vol. 2, pp. 1329–1332, Oct. 2004.
40. T. Wiegand, G. J. Sullivan, J. Reichel, H. Schwarz, and M.Wien, "Joint Draft ITU-T Rec. H.264|ISO/IEC 14496-10 / Amd.3 Scalable video coding", *ISO/IEC JTC1/SC29/WG11 and ITU-T SG16 Q.6 Doc. JVT-X201*, Doc. JVT-X201, Geneva, Switzerland, July 2007.

41. H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the Scalable Video Coding Extension of the H.264/AVC Standard", *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no 9, pp. 1103–1120, Sept. 2007.
42. E. Itakura, S. Futemma, W. Guijin, and K. Yamane, "JPEG2000 based real-time scalable video communication system over the Internet", *Second IEEE Consumer Communications and Networking Conference, CCNC'05*, pp. 539–543, Jan. 2005.
43. S. Wenger, Y.-K. Wang, T. Schierl, and A. Eleftheriadis, "Internet-Draft RTP Payload Format for SVC Video", *draft-ietf-avt-rtp-svc-14.txt*, September 2008.
44. A. Smolic, K. Mueller, N. Stefanoski, J. Ostermann, A. Gotchev, G.B. Akar, G. Triantafyllidis, and A. Koz, "Coding Algorithms for 3DTV – A Survey", *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 11, pp. 1606–1621, Nov. 2007.
45. P. Merkle, K. Muller, A. Smolic, and T. Wiegand, "Statistical Evaluation of Spatio-Temporal Prediction for Multi-View Video Coding", *2*nd *Workshop on Immersive Communication and Broadcast Systems (ICOB'05)*, Berlin, Germany, October 2005.
46. P. Merkle, K. Muller, A. Smolic, and T. Wiegand, "Efficient Compression of Multi-View Video Exploiting Inter-View Dependencies Based on H.264/MPEG4-AVC", *IEEE International Conference on Multimedia and Expo (ICME'06)*, Toronto, Ontario, Canada, July 2006.
47. E. Ekmekcioglu, S.T. Worrall, and A.M. Kondoz, "Low-delay random view access in multi-view coding using a bit-rate adaptive down-sampling approach", *IEEE International Conference on Multimedia and Expo*, pp. 745–748, April 2008.
48. B. Balasubramaniyam, E. Edirisinghe, and H. Bez, "An extended H.264 CODEC for stereoscopic video coding", *Proceedings of SPIE – The International Society for Optical Engineering*, pp. 116–126, 2005.
49. A. Redert, E. Hendriks, and J. Biemond, "Correspondence Estimation in Image Pairs", *IEEE Signal Processing Magazine*, vol. 16, no. 3, pp. 29–46, May 1999.
50. M.S. Moellenho, and M.W. Maier, "DCT Transform Coding of Stereo Images for Multimedia Applications", *IEEE Transactions on Industrial Electronics*, vol. 45, no. 1, pp. 38–43, Feb. 1998.
51. M.G. Strintzis, and S. Malassiotis, "Object-based Coding of Stereoscopic and 3D Image Sequences", *IEEE Signal Processing Magazine*, vol. 16, no. 3, pp. 14–28, May 1999.
52. A. Aksay, C. Bilen, E. Kurutepe, T. Ozcelebi, G.B. Akar, M.R. Civanlar, and A.M. Tekalp, "Temporal and Spatial Scaling for Stereoscopic Video Compression", *IEEE 14th Eur. Signal Process. Conf. EUSIPCO 2006*, Florence, Italy, Sept. 2006.
53. L. Christodoulou, L. Mayron, H. Kalva, O. Marques, and B. Furht, "Design and Evaluation of 3-D Video System Based on H.264 View Coding", *International Workshop on Network and Operating Systems Support for Digital Audio and Video (NOSSDAV 2006)*, Newport, Rhode Island, May, 2006.
54. MPEG-2 Multiview profile, ISO/IEC 13818-2, AMD 3, Doc. no. N1366, 1996.

55. C. Fehn and A. Smolic, "Study of some MPEG Tools Related to 3-D-Video", *ISO/IEC JTC1/SC29/WG11 M8423*, Doc. no. M8423, Fairfax, May 2002.
56. H. Jia, W. Gao, and Y. Lu, "Stereoscopic Video Coding Based on Global Displacement Compensated Prediction", *IEEE ICICS-PCM 2003*, pp. 61–65, Singapore, Dec. 2003.
57. Generic Coding of Audio-Visual Objects – Part 2: Visual, ISO/IEC 14496-2, Doc. no. N4350, 2001.
58. M. Bourges-Sevenier, and E.S. Jang, "An introduction to the MPEG-4 animation framework eXtension", *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 7, pp. 928–936, July 2004.
59. C. Fehn, K. Schr, P. Kau, and A. Smolic, Coding Results for EE4 in MPEG 3DAV, *ISO/IEC JTC1/SC29/WG11 MPEG02/M9561* , Pattaya, March 2003.
60. G.B. Akar, A.M. Tekalp, C. Fehn, and M.R. Civanlar, "Transport Methods in 3DTV – A Survey", *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 11, pp. 1622–1630, Nov. 2007.
61. International Broadcasting Convention (IBC) 2008, "Transatlantic High Definition Stereoscopic 3D broadcast", Amsterdam, Netherland, September 2008. Available: http://eandt.theiet.org/magazine/2008/17/game-o-leapfrog-0817.cfm.
62. S. Pehlivan, A. Aksay, c. Bilen, G.B. Akar, and M.R. Civanlar, "End-to-End Stereoscopic Video Streaming System" *IEEE International Conference on Multimedia and Expo*, pp. 2169–2172, July 2006.

63. P. Kauff, O. Schreer, and R. Tanger, "Virtual team user environments – a mixed reality approach for immersive tele-collaboration" *In Proceedings of International Workshop Immersive Telepresense (ITP2002)*, pp. 1–4, Juan Les Pins, France, December, 2002.
64. R.M. Hayes, 3-D Movies: A History and Filmography of Stereoscopic Cinema, New York: McFarland, 1998.
65. R. Sand, "3D -TV – A review of recent and current developments", *in. Proc. IEE Colloq. Stereoscopic Television*, pp. 1–4, London, UK, Oct. 1992.
66. I. Yuyama, and M. Okui, "Stereoscopic HDTV", *in Three-Dimensional Television, Video, and Display Technologies*, B. Javidi and F. Okano, Eds. New York: Springer-Verlag, pp. 3–34, 2002.
67. N.H. Hur, C.H. Ahn, and C.T. Ahn, "Experimental Service of 3DTV Broadcasting Relay in Korea", *Three-Dimensional TV, Video and Display, Proc. SPIE 4864*, pp. 1–13, 2002.
68. P. Harman, "An architecture for digital 3-D broadcasting", *In Proc. SPIE Stereoscopic displays and virtual reality systems VI*, vol. 3639, pp. 254–259, Sanjose CA, January 1999.
69. J. Flack, P. Harman, and S. Fox, "Low bandwidth stereoscopic image encoding and transmission", *In Proc. SPIE Stereoscopic displays and virtual reality systems X*, vol. 5006, pp. 206–214, 2003.
70. J. van der Meer, and A. Bourge, "Carriage of Auxiliary Video Data", *ISO/IECJTC 1/SC 29/WG 11. FPDAM of ISO/IEC 13818-1:200X/AMD 2, WG 11*, Doc. N8094, Jul. 2006.
71. A.M. Tekalp, E. Kurutepe, and M.R. Civanlar, "3DTV over IP", *IEEE Signal Processing Magazine*, vol. 24, no. 6, pp. 77–87, Nov. 2007.
72. E. Kohler, M. Handley, and S. Floyd, "Datagram congestion control protocol (DCCP)", *Internet Engineering Task Force RFC 4340*, Mar. 2006.
73. S. Floyd, E. Kohler, and J. Padhye, "Profile for DCCP congestion control ID 3: TCP-friendly rate control (TFRC)", *Internet Engineering Task Force RFC 4342*, Mar. 2006.
74. N. Ozbek and A.M. Tekalp, "Content-aware bit allocation in scalable multiview video coding", *in Proc. Int. Workshop Multimedia Content Representat., Classificat. Security (MCRS)*, vol. 4105, pp. 691–698, 2006.
75. E. Kurutepe, M.R. Civanlar, and A.M. Tekalp, "A receiver-driven multicasting framework for 3DTV transmission", *presented at the 13th Eur. Signal Process. Conf. (EURASIP)*, Antalya, Turkey, Sep. 2005.
76. E. Kurutepe, M.R. Civanlar, and A.M. Tekalp, "Interactive transport of multiview videos for 3DTV applications", *J. Zhejiang Univ. Science A*, vol. 7, no. 5, pp. 830–836, 2006.
77. M. Bystrom, J.W. Modestino, "Combined source-channel coding schemes for video transmission over an additive white Gaussian noise channel", *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 6, pp. 880–890, Jun 2000.
78. B. Kamolrat, W.A.C. Fernando, M. Mrak, A. Kondoz, "Joint source and channel coding for 3-D video with depth image – based rendering", *IEEE Transactions on Consumer Electronics*, vol. 54, no. 2, pp. 887–894, May 2008.

79. S. Argyropoulos, A.S. Tan, N. Thomos, E. Arikan, and M.G. Strintzis, "Robust transmission of multiview video streams using flexible macroblock ordering and systematic LT codes", *presented at the 3DTVCON*, Kos Island, Greece, May 2007.
80. A.S. Tan, A. Aksay, C. Bilen, G.B. Akar, and E. Arikan, "Error resilient layered stereoscopic video streaming", *presented at the 3DTVCON*, Kos Island, Greece, May 2007.
81. H.A. Karim, C.T.E.R. Hewage, S.T. Worrall, and A.M. Kondoz, "Scalable multiple description video coding for stereoscopic 3-D", *IEEE Transactions on Consumer Electronics*, vol. 54, no. 2, pp. 745–752, May 2008.
82. S. Belfiore, M. Grangetto, E. Magli and G. Olmo, "Concealment of whole frame losses for wireless low bit-rate video based on multi-frame optical flow estimation", *IEEE Transactions on Multimedia*, vol. 7, no. 2, pp. 316–329, 2005.
83. P. Baccichet, A. Chimienti, D. Bagni, F.S. Rovati, and L. Pe-zoni, "Frame Concealment for H.264/AVC decoders", *IEEE Transactions on Consumer Electronics*, vol. 51, no. 1, pp. 227–233, February 2005.
84. S. Knorr, C. Clemens, M. Kunter, and T. Sikora, "Robust concealment for erroneous block bursts in stereoscopic images", *In Proc. 2*nd *International Symposium 3-D Data Process., Visual., Transmission (3DPVT)*, pp. 820–827, 2004.
85. X. Xiang, D. Zhao, Q. Wang, X. Ji and W. Gao, "A novel error concealment method for stereoscopic video coding", *International conference on Image processing(ICIP-2007)*, pp. 101–104, 2007.
86. C. Bilen, A. Aksay and Gozde Akar, "Two novel methods for full frame loss concealment in stereo video", *In Proceedings of 26*th *Picture coding Symposium (PCS'07)*, 2007.
87. L. Pang, M. Yu, G. Jiang, Z. Jiang, and F. Li, "An Approach to Error Concealment for Entire Right Frame Loss in Stereoscopic Video Transmission", *International Conference on Computational Intelligence and Security,* vol. 2, pp. 1665–1670, Nov. 2006.
88. P. Benzie, J. Watson, P. Surman, I. Rakkolainen, K. Hopf, H. Urey, V. Sainov, and C. von Kopylow, "A Survey of 3DTV Displays: Techniques and Technologies", *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 11, pp. 1647–1658, November 2007.
89. B.T. Schowengerdt, and E.J. Seibel, "True 3-D scanned voxel displays using single and multiple light sources", *Journal Soc. Inf. Display*, vol. 14, no. 2, pp. 135–143, 2006.
90. D. Bahr, K. Langhans, M. Gerken, C. Vogt, D. Bezecny, and D. Homann, "Felix: A volumetric 3-D laser display", *in Proc. SPIE Projection Display. II*, vol. 2650, pp. 265–273, San Jose, CA, 1996.
91. S. DiVerdi, I. Rakkolainen, T. Hollerer, and A. Olwal, "A novel walkthrough 3-D display", *in Proc. SPIE, Stereoscop. Displ. Virt. Real. Syst. XIII*, vol. 6055, pp. 1–10, 2007.
92. L.H. Enloe, J.A. Murphy, and C. B. Rubinstein, "Hologram transmission via television", *Bell Syst. Tech. Journal*, vol. 45, no. 2, pp. 225–339, 1966.
93. P.M. Hubel, "Recent advances in colour reflection holography", *in Proc. SPIE, Practical Hologr. V, S.A. Benton, Ed.*, vol. 1461, pp. 167–174, 1991.
94. T.-C. Poon, "Three-dimensional television system using optical scanning holography and spatial light modulator", *Journal of Inf. Displ.*, vol. 3, no.12, 2002.

95. T.-C. Poon, "Recent progress in optical scanning holography", *Journal of Holography and Speckle*, vol. 1, pp. 6–25, 2004.
96. I.P. Howard and B. J. Rogers, *Seeing in Depth: Depth Perception*, vol. 2, I. Porteous Publishing, Toronto, Canada: 2002.
97. N.A. Dodgson, "Autostereoscopic 3-D displays", *Computer*, vol. 38, no. 8, pp. 31–36, 2005.
98. Sharp, "Sharp 3-D", Available: http://www.sharp.eu/cps/rde/xchg/eu/hs.xsl/-/html/aquos-3d-technology.htm, Sep 2014.
99. Philips, "Philips 3-D solutions", Available: http://www.philips.co.uk/c-p/BDL5071VS_00/signage-solutions-50-inch-edge-led-backlight-ultra-hd-autostereoscopic-3d/overview, 2014.
100. W. IJsselsteijn, H. de Ridder, and J. Vliegen, "Subjective evaluation of stereoscopic images: Effects of camera parameters and display duration", *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 10, no. 2, pp. 225–233, 2000.
101. J. Freeman and S. Avons, "Focus group exploration of presence through advanced broadcast services", *Proceedings of the SPIE 3959*, pp. 530–539, 2000.
102. J.A.J. Roufs, "Perceptual image quality: Concept and measurement", *Philips Journal of Research 47*, pp. 35–62, 1992.
103. International Telecommunication Union/ITU Radio communication Sector, "Subjective Assessment of Stereoscopic Television Pictures", *ITU-R BT.1438*, Jan 2000.
104. International Telecommunication Union/ITU Radio communication Sector, "Methodology for the subjective assessment of the quality of television pictures", *ITU-R BT.500-11*, Jan 2002.

105. P. Seuntiens, L. Meesters, and W. Ijsselsteijn, "Perceived quality of compressed stereoscopic images: Effects of symmetric and asymmetric JPEG coding and camera separation", *ACM Transactions on Applied Perception (TAP)*, vol. 3, no. 2, pp. 95–109, April 2008.
106. A. Aksay, C. Bilen, and G. Bozdagi, "Subjective evaluation of effects of spectral and spatial redundancy reduction on stereo images", *13th European Signal Processing Conference*, Antalya, September 2005.
107. P. Seuntiens, L. Meesters, and W. IJsselsteijn, "Perceptual evaluation of JPEG coded stereoscopic images", *Proceedings of SPIE*, vol. 5006, pp. 215–226, 2003.
108. H.A. Karim, S. Worrall, A.H. Sadka, A. M. Kondoz, "3-D video compression using MPEG4-multiple auxiliary component (MPEG4-MAC)", *IEE 2nd International Conference on Visual Information Engineering (VIE2005)*, April 2005.
109. J. Kim, Y. Kim, and K. Sohn, "Stereoscopic video coding and disparity estimation for low bitrate applications based on MPEG-4 multiple auxiliary components", *Image Communication*, vol. 23, no. 6, pp. 405–416, July 2008.
110. M. Martini, C. Chen, Z. Chen, T. Dagiuklas, L. Sun, and X. Zhu, "Guest editorial: QoE-aware wireless multimedia systems," *IEEE Journal on Selected Areas in Communications*, vol. 30, no. 7, pp. 1153–1156, 2012.
111. P. Le-Callet, S. Moeller, and A. Perkis, "Qualinet white paper on definitions of quality of experience, version 1.1," *in European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003)*, June 2012.
112. K. Wang, M. Barkowsky, K. Brunnstrom, M. Sjostrom, R. Cousseau, and P. Le-Callet, "Perceived 3-D TV transmission quality assessment: Multi-laboratory results using absolute category rating on quality of experience scale," *IEEE Transactions on Broadcasting*, vol. 58, no. 4, pp. 544–557, 2012.
113. M. Lambooij, M. Fortuin, I. Heynderickx, and W. IJsselsteijn, "Visual discomfort and visual fatigue of stereoscopic displays: A review," *Journal of Imaging Science and Technology*, vol. 53, no. 3, pp. 30 201–1 – 30 201–14, May 2009.
114. G.J. Sullivan, J. Ohm, W.J. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard", IEEE Transactions on Circuits and Systems for Video Technology, vol. 22, no. 12, pp. 1649–1668, 2012.
115. L. Meesters, W. IJsselsteijn, and P. Seuntiens, "A survey of perceptual evaluations and requirements of three-dimensional TV," IEEE Transactions on Circuits and Systems for Video Technology, vol. 14, no. 3, pp. 381–391, March 2004.
116. M. Martini, C. Chen, Z. Chen, T. Dagiuklas, L. Sun, and X. Zhu, "Guest editorial: QoE-aware wireless multimedia systems," IEEE Journal on Selected Areas in Communications, vol. 30, no. 7, pp. 1153–1156, 2012.
117. P. Le-Callet, S. Moeller, and A. Perkis, "Qualinet white paper on definitions of quality of experience, version 1.1," in European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003), June 2012.

118. K. Wang, M. Barkowsky, K. Brunnstrom, M. Sjostrom, R. Cousseau, and P. Le-Callet, "Perceived 3D TV transmission quality assessment: Multilaboratory results using absolute category rating on quality of experience scale," IEEE Transactions on Broadcasting, vol. 58, no. 4, 2012.

119. J. Cutting and P. Vishton, Perceiving layout and knowing distances: The integration, relative potency, and contextual use of different information about depth, in Perception of Space and Motion, edited by W. Epstein and S. Rogers, pp. 69–117, 1995.

120. M. Lambooij, M. Fortuin, I. Heynderickx, and W. IJsselsteijn, "Visual discomfort and visual fatigue of stereoscopic displays: A review," Journal of Imaging Science and Technology, vol. 53, no. 3, pp. 30 201–1– 30 201–14, May 2009.

121. P. Seuntiens, L. Meesters, and W. Ijsselsteijn, "Perceived quality of compressed stereoscopic images: Effects of symmetric and asymmetric JPEG coding and camera separation," ACM Transactions on Applied Perception (TAP), vol. 3, no. 2, pp. 95–109, 2006.

122. A. Benoit, P. Le-Callet, P. Campisi, and R. Cousseau, "Quality assessment of stereoscopic images," EURASIP journal on image and video processing, vol. 2008, 2008.

123. C.T.E.R. Hewage, S.T. Worrall, S. Dogan, S. Villette, and A.M. Kondoz, "Quality evaluation of color plus depth map-based stereoscopic video," IEEE Journal of Selected Topics in Signal Processing, vol. 3, no. 2, pp. 304–318, 2009.

124. J. Seo, X. Liu, D. Kim, and K. Sohn, "An objective video quality metric for compressed stereoscopic video," Circuits, Systems, and Signal Processing, pp. 1–19, 2012.

125. M. Martini, M. Mazzotti, C. Lamy-Bergot, J. Huusko, and P. Amon, “Content adaptive network aware joint optimization of wireless video transmission,” IEEE Communications Magazine, vol. 45, no. 1, pp. 84–90, 2007.

126. C.T.E.R. Hewage, and M.G. Martini, “Edge-based reduced-reference quality metric for 3-D video compression and transmission,” IEEE Journal of Selected Topics in Signal Processing, vol. 6, no. 5, pp. 471–482, 2012.

127. Z. Sazzad, S. Yamanaka, Y. Kawayokeita, and Y. Horita, “Stereoscopic image quality prediction,” in Proc. International Workshop on Quality of Multimedia Experience (QoMEX), 2009, pp. 180–185.

128. A. Maalouf and M. Larabi, “CYCLOP: a stereo color image quality assessment metric,” in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2011, pp. 1161–1164.

129. M. Solh and G. AlRegib, “A no-reference quality measure for DIBR based 3D videos,” in Proc. IEEE International Conference on Multimedia and Expo (ICME), 2011, pp. 1–6.

130. C.T.E.R. Hewage, S.T. Worrall, S. Dogan, H. Kodikaraarachchi, and A.M. Kondoz, “Stereoscopic TV over IP”. *In: 4th European Conference on Visual Media Production;* 27–28 Nov 2007, London, U.K.

131. C.T.E.R. Hewage, S.T. Worrall, S. Dogan, and A.M. Kondoz, “A Novel frame concealment method for depth maps using corresponding colour motion vectors”. *In: 3DTV Conference: The True Vision – Capture, Transmission and Display of 3D Video;* 28-30 May 2008, Istanbul, Turkey.

132. M.G. Martini, C.T.E.R. Hewage, S. Nasir, and S.T. Worrall, “Prioritized 3D video distribution over IEEE 802.11e”, *In: Future Network & Mobile Summit;* 16 June–18 June 2010, Florence, Italy.

133. C.T.E.R. Hewage, Z. Ahmad, S.T. Worrall, S. Dogan, W.A.C. Fernando, and A.M. Kondoz, “Unequal Error Protection for backward compatible 3-D video transmission over WiMAX”, *In: IEEE International Symposium on Circuits and Systems;* 24–27 May 2009, Taipei, Taiwan.

134. B. Randolph, “A neural theory of binocular rivalry”, *Psychological review*, vol. 96, no. 1, pp. 145–167, 1989.